CANOPY ARCHITECTURE AND PERFORMANCE MODELING IN *Sorghum bicolor*

A Dissertation

by

SANDRA KHAVI TRUONG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	John Mullet
Co-Chair of Committee,	Peter Howard
Committee Members,	Michael Polymenis
	Lanying Zeng
Chair of Intercollegiate Faculty,	Dorothy Shippen

May 2017

Major Subject: Genetics

ABSTRACT

The global population is increasing in size and economic affluence, and sustainable intensification of agriculture is necessary to meet projected fuel and food demands without increasing input of limited resources like water. One approach to sustainably intensify agriculture is to improve the productivity of food and biofuel crops by increasing photosynthetic efficiency through modulation of canopy features. Sorghum is the world's fifth leading cereal crop and a promising bioenergy crop, and represents a useful target for crop improvement. Genetically engineering sorghum canopy architecture ideotypes for target production environments is facilitated by combining mathematical modeling, *in silico* simulation, field experimentation, and genetic analyses to determine optimal combinations of traits and the genetic networks regulating those traits, and this dissertation describes work towards this end. First, a model to improve calculations of genetic linkage in plant recombinant inbred line populations was developed so that regions of the genome associated with phenotypic variation can be more accurately defined. Second, functional-structural plant modeling was combined with field experimentation and genetic analyses to quantify the influence of leaf angle, a canopy architecture trait, on plant productivity, as well as to identify genomic loci regulating leaf angle. Third, a bioenergy sorghum crop model was parameterized to predict production of lignocellulosic biomass; this model was extended to explore how different canopy transpiration strategies influenced the end productivity of the crop and identify transpiration strategies yielding optimal water use. Continued progress in this work will provide ways to evaluate germplasm performance *in silico* and guide field testing for crop improvement.

ACKNOWLEDGEMENTS

The work described in this document was made possible by the contributions of multiple individuals, organizations, software development teams, and funding agencies.

I thank my Ph.D. committee co-chairs, Professor John Mullet and Professor Peter Howard, and the members of the committee, Professor Michael Polymenis and Professor Lanying Zeng, for their guidance and support during this work. I thank all current and former members of the John Mullet lab for their assistance and advice, particularly Ryan McCormick, Josie Hilley, and Daryl Morishige, Ph.D.. I thank the current and former members of the William Rooney lab for their assistance and collaboration on the work presented here. I thank the Texas A&M Institute for Genome Sciences and Society for maintaining the TIGSS-HPC cluster used for many of the analyses reported herein. I thank the following funding sources: Department of Energy Great Lakes Bioenergy Research Center (BER DE-FC02-07ER64494), Department of Energy ARPA-E TERRA project (DE-AR0000596), and the Perry Adkisson Chair in Agricultural Biology. The author also thanks external collaborators at the Department of Energy Joint Genome Institute, the Hudson Alpha Institute for Biotechnology, the Robotics Instititute at Carnegie Mellon University, and the National Robotics Engineering Center.

I thank the developers and communities surrounding the numerous open-source software projects used for their useful software, helpful advice, forum discussions, and for making their source code available. These projects and organizations include: Linux, the GNU Operating System, Canonical, the Standard C++ Foundation, Python software foundation, the R core team, PCL, Open CV, Meshlab, VTK,

GATK, IGV, Pandas, Matplotlib, SciPy, seaborn, R/qtl, OpenAlea, VTK library, ParaView, URGI for the REPET package, SunGridEngine, the HTSeq library, Foswiki, Code::Blocks, TexWorks, SequenceServer, GNU Parallel, and the Document Foundation for LibreOffice, and others that may have been used but were failed to be remembered. The author also thanks the Stack Exchange community for their constructive technical insights.

And finally, I thank the journal editors and anonymous reviewers that provided constructive feedback on the manuscripts that are reproduced herein, as well as the journals themselves for permitting reproduction.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor John Mullet of the Department of Biochemistry and Biophysics, Professor Peter Howard of the Department of Mathematics, Professor Michael Polymenis of the Department of Biochemistry and Biophysics, and Professor Lanying Zeng of the Department of Biochemistry and Biophysics.

Work for the dissertation was completed by Sandra Truong in collaboration with Ryan McCormick, Daryl Morishige, William Rooney and John Mullet, co-authors on the manuscripts reproduced herein.

Funding Sources

This work was supported by the following funding sources: Department of Energy Great Lakes Bioenergy Research Center (BER DE-FC02-07ER64494), Department of Energy ARPA-E TERRA project (DE-AR0000596), and the Perry Adkisson Chair in Agricultural Biology.

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGEMENTS iii

CONTRIBUTORS AND FUNDING SOURCES v

TABLE OF CONTENTS vi

LIST OF FIGURES ix

LIST OF TABLES xi

1. INTRODUCTION 1

2. RESOLUTION OF GENETIC MAP EXPANSION CAUSED BY EXCESS HETEROZYGOSITY IN PLANT RECOMBINANT INBRED POPULATIONS 6

2.1 Overview 6
2.2 Introduction 6
2.3 Results 10
2.3.1 Excess heterozygosity generally causes overestimation of recombination frequencies 10
2.3.2 Incorporation of a heterozygosity term into the genetic model shrinks a sorghum genetic map 12
2.4 Discussion 14
2.5 Materials and methods 18
2.5.1 Derivation and implementation of genetic model 18
2.5.2 Plant materials and genotyping 23
2.5.3 Genetic map construction 25

3. HARNESSING GENETIC VARIATION IN LEAF ANGLE TO INCREASE PRODUCTIVITY OF *Sorghum bicolor* 27

3.1 Overview 27
3.2 Introduction 28
3.3 Methods 30

3.3.1 Genetic basis of leaf inclination angle . . . 30
3.3.2 Phenotyping leaf inclination angle . . . 31
3.3.3 Stepwise multiple-QTL mapping and heritability estimates . . 32
3.3.4 *dw3* gene identification . . . 33
3.3.5 Calculation of the light extinction coefficient, k . . . 34
3.3.6 Virtual sorghum canopies . . . 34
3.3.7 Field experimental sorghum canopies . . . 35
3.3.8 Estimating potential biomass gain . . . 35
3.4 Results . . . 36
3.4.1 Genetic regulation of sorghum leaf angle . . . 36
3.4.2 *dwarf-3* regulates leaf angle in sorghum. . . . 38
3.4.3 Leaf angle affects vertical light distribution in sorghum canopies. 38
3.5 Discussion . . . 45

4. BIOENERGY SORGHUM CROP MODEL PREDICTS VPD-LIMITED TRANSPIRATION TRAITS ENHANCE BIOMASS YIELD IN WATER-LIMITED ENVIRONMENTS . . . 49

4.1 Overview . . . 49
4.2 Introduction . . . 50
4.3 Methods . . . 53
4.3.1 Crop model simulations for energy sorghum in APSIM . . . 53
4.3.2 Adapting APSIM sorghum module for the VPD-limited transpiration trait . . . 56
4.3.3 Data and code accessibility . . . 60
4.4 Results . . . 60
4.4.1 Modeling bioenergy sorghum in APSIM . . . 60
4.4.2 Modeling VPD-limited transpiration trait . . . 63
4.5 Discussion . . . 72

5. CONCLUSION . . . 77

REFERENCES . . . 79

APPENDIX A. SUPPLEMENTAL MATERIAL FOR RESOLUTION OF GENETIC MAP EXPANSION CAUSED BY EXCESS HETEROZYGOSITY IN PLANT RECOMBINANT INBRED POPULATIONS . . . 97

A.1 Solving for the general solution, p_{F_t}, in MATLAB . . . 97
A.2 Parameterization of the heterozygosity term . . . 102
A.2.1 A global heterozygosity term . . . 103
A.2.2 Local heterozygosity term for each linkage group, $h_{\text{linkage group \#}}$ 104

A.2.3 The derivation for a local heterozygosity term for each marker, h_{marker} . . . 104
A.3 Genetic map estimations of simulated datasets with excess heterozygosity . . . 108
A.3.1 Tight double recombinations . . . 109
A.3.2 Error probability in the hidden Markov model . . . 110
A.3.3 F7 versus fixed RIL ($t \to \infty$) . . . 110

APPENDIX B. SUPPLEMENTARY MATERIAL FOR HARNESSING GENETIC VARIATION IN LEAF ANGLE TO INCREASE PRODUCTIVITY OF *Sorghum bicolor* . . . 115

B.1 Extended materials and methods . . . 115
B.1.1 Calculation of the light extinction coefficient, k. . . . 115
B.1.2 Virtual sorghum canopies. . . . 117
B.1.3 Field experimental sorghum canopies. . . . 118
B.2 Files SI . . . 121
B.2.1 Additional supporting folders and files . . . 121

APPENDIX C. SUPPLEMENTARY MATERIAL FOR BIOENERGY SORGHUM CROP MODEL PREDICTS VPD-LIMITED TRANSPIRATION TRAITS ENHANCE BIOMASS YIELD IN WATER-LIMITED ENVIRONMENTS 126

C.1 VPD-limited transpiration in simulated water-environments . . . 126

LIST OF FIGURES

FIGURE Page

2.1 Estimated recombination frequencies, $\hat{r}$, under excess heterozygosity and Mendelian models. 11

2.2 Excess heterozygosity in a sorghum mapping population. 13

2.3 Accounting for excess heterozygosity shrinks the sorghum genetic map. 16

3.1 *dw3* regulates leaf inclination angle. 40

3.2 Leaf inclination angle affects light distribution in sorghum canopies. . 42

4.1 Growth simulations and field data of energy sorghum in unlimited, limited, and no irrigation regimes. 62

4.2 Seasonal fluctuations of water and VPD and their relationship to sorghum transpiration rate and efficiency. 64

4.3 Performance of VPD-limited transpiration determined by drought condition. 67

4.4 Energy sorghum biomass accumulation from 2000-2014 cropping seasons simulated in College Station, TX, with water supply coming only from rainfall. 70

4.5 Landscapes of energy sorghum productivity for varying m_1, m_2 and vpd_{BP} within the observed phenotypic parameters. 75

4.6 Theoretical energy sorghum productivity landscape. 76

A.1 Estimated recombination fractions, $\hat{r}$, of excess heterozygosity versus Mendelian expectations for $t = 3$. 99

A.2 Heterozygosity landscape. 112

A.3 Heterozygosity distribution in sorghum F7 mapping population. . . . 113

A.4 Heterozygosity on chromosome 2. 114

A.5 Screenshot of spreadsheet containing map estimation results for the sorghum mapping population and simulated data with different models and methods. 114

B.1 *dw3* regulates leaf inclination angle. 122

B.2 Lift for LP-80 PAR meter. 123

B.3 *dwarf-3* genotypes of the bi-parental mapping populations. 123

B.4 R07018 x R07020 RIL 63 & RIL 73 phenotypes. 124

B.5 Leaf inclination angle regulates light distribution in canopies. 125

C.1 Growth simulations of energy sorghum agree with field data for 2009 under the limited irrigation regime. 129

C.2 Energy sorghum biomass accumulation from 2000-2014 cropping seasons simulated in College Station, TX, with water supply regimes of unlimited irrigation and only rainfed. 130

C.3 Predictions of biomass accumulation by bioenergy sorghum with VPD-limited transpiration traits grown in rain fed cropping conditions of College Station, TX, 2000 2014 against days after sowing (DAS). . . 131

C.4 Performance of VPD-limited transpiration in low and high yielding rain-fed cropping environments. 132

C.5 Performance of VPD-limited transpiration determined by canopy maintenance. 133

C.6 Predictions of biomass accumulation by bioenergy sorghum with VPD-limited transpiration traits grown in simulated terminal drought and well-watered environments. 134

LIST OF TABLES

TABLE Page

2.1 Genetic maps estimated from the BTx623 x IS3620c sorghum recombinant inbred mapping population. 15

3.1 Experimental crosses and phenotyping metadata. 37

3.2 Trait heritability and QTL of leaf inclination angle. 39

4.1 Parameter values for modeling phenology, canopy development, and growth in the energy sorghum crop model used in this study. 55

4.2 Descriptions of variables in the VPD-limited transpiration model modified within the APSIM sorghum module. 59

4.3 Descriptions of variables in the VPD-limited transpiration model modified within the APSIM sorghum module. 65

A.1 Reports of deviation from expected heterozygosity maintained per generation. 97

A.2 Reported genetic map sizes for the sorghum BTx623 x IS3620c RIL population used in this study. 98

B.1 Light distribution in canopies of RIL 63 and RIL 73. 120

C.1 Environmental data and end biomass of energy sorghum that vary in their m_2 simulations in the rain-fed cropping environments of College Station, TX. 135

1. INTRODUCTION

Agricultural demands imposed by global population growth and expansion of the middle class is projected to require a doubling of food production by the year 2050, but yield gains of major agricultural crops have plateaued (Godfray *et al.*, 2010; Alexandratos and Bruinsma, 2012; Foley *et al.*, 2011; Tilman *et al.*, 2011). For cereal crops, the projected and current yield trends of 0.9% yield gain per year since 2007 would be sufficient to meet projected food demands (Godfray *et al.*, 2010; Porter *et al.*, 2010; Mueller *et al.*, 2012; Alexandratos and Bruinsma, 2012). In addition to greater caloric demands, this more affluent population will also consume for fuel leading to increased energy demands. However, to meet projected demands without any further agricultural innovation would require a scaling of production that would require greater input of land, fertilizer, pesticide, and water. Given that fertile land and fresh water are already limited resources, further expansion of resource intensive and destructive agricultural practices is undesirable. In light of sustaining the global environment, improving crop yields must prioritize research that expands agricultural output without further depleting natural resources (Foley *et al.*, 2011; Mueller *et al.*, 2012). Multiple research foci are developing new crop varieties that offer higher yields and use less water, fertilizer, and pesticides including studies of drought tolerance, biofertilizing rhizobacteria, and resistance to pests. This dissertation focuses on developing and implementing genetic and mathematical modeling tools to improve crop yields for both energy and food by optimizing crop canopy interception of solar radiation and water use.

In particular, this work is focused on the improvement of the crop plant *Sorghum bicolor*, the fifth most widely grown cereal crop that is also grown for biofuel, sugar,

and forage. Sorghum utilizes C4 photosynthesis which outperform C3 photosynthetic grasses in hot and dry environments (Mullet *et al.*, 2014). Sorghum has a relatively small sequenced genome (Paterson *et al.*, 2009) and is amenable to controlled crosses and hybrid breeding (Rooney *et al.*, 2007). The research described in this dissertation will aid in the design of more productive sorghum, and progress in sorghum can be easily translated to other related agricultural grasses, including maize and rice.

One approach to improve sorghum performance is to genetically engineer desirable traits for target environments. In this way, we can dictate the underlying regulation of plant phenotypes. In order to design or select genetic ideotypes, we rely on foundational knowledge of the genetic networks regulating target phenotypes and how these control development with respect to genetic and environmental perturbations. Identification of loci in the genome that are associated with phenotypic regulation is typically made by finding a correlation between genotypic variation with phenotypic variation. The identification of such loci can be improved by taking advantage of information on how tightly linked regions of the genome are with one another, enabling advanced association mapping analyses like interval mapping. Interval mapping and related analyses require accurate estimates of genetic linkage (i.e. an accurate genetic map), and experimental plant populations often display skewed inheritance schemes that cause linkage mapping estimates to be inaccurate when estimated using traditional genetic models. The work in chapter 2 models and implements a method to account for differential zygotic viability of heterozygous loci in the construction of a genetic map, and the method allows for the retainment and use of more genetic information. The particular population motivating the development of the model and its genetic map has been subsequently used in to identify genomic loci underlying the regulation of important agronomic traits (Truong *et al.*, 2015; Hilley *et al.*, 2016; Mccormick *et al.*, 2016). In particular, the population has

been used to examine the genetic basis of traits influencing canopy architecture and in consequence on environment interaction (i.e. solar radiation).

Crops intercept photosynthetically active radiation (PAR) from the sun to convert atmospheric carbon into chemical energy to fuel plant processes like grain production, crop canopy architecture affects the distribution and interception of solar radiation. Canopies that optimize PAR interception can potentially increase their photosynthetic efficiency and, in turn, yield. On a per leaf basis, more optimal solar radiation distribution is thought to play a role in mitigating heat stress induced by excess infrared radiation as well as maximizing carbon gain through optimal interception of PAR for canopy photosynthesis (Zhu *et al.*, 2010; van Zanten *et al.*, 2010; Song *et al.*, 2013). In certain cropping environments, leaves at the top of the canopy are often oversaturated in the midday hours (Long *et al.*, 2006; Zhu *et al.*, 2010; Drewry *et al.*, 2014). Interception of PAR in excess of that needed to saturate photosynthetic electron transport leads to a decline in energy conversion efficiency and, in some instances, to photo-inhibition (Zhu *et al.*, 2010; Long *et al.*, 2006; Murchie *et al.*, 1999). Interception of excess infrared radiation (IR) heats the leaves which increases vapor pressure deficit (VPD) and transpiration, resulting in reduced water use efficiency (Nobel, 2005; Ort and Long, 2014). In contrast to leaves at the top of the canopy, leaves at the bottom of the canopy are often undersaturated for solar radiation. Insufficient solar radiation can lead to shade-induced senescence, and the resulting decrease in green leaf area reduces accumulation of nitrogen, a possible rate-limiting factor during grain filling in modern high-yielding cultivars (Sinclair and Sheehy, 1999; Hammer *et al.*, 2009; van Zanten *et al.*, 2010). The work in chapter 3 examines the impact of canopy architecture on light interception in canopies *in silico* and in field experimentation of the sorghum canopy, identifies quantitative trait loci (QTL) that regulate leaf angle, a particular crop canopy architecture trait,

in mapping populations, and identifies a gene (*Dwarf3*) that changes leaf angle up to 30 degrees in sorghum.

Whereas solar radiation is often available in excess, water is a major limiting environmental input for agriculture, and resources that can accelerate the identification of, or prioritization of, traits for water-efficient crops will be critical to increase crop productivity. Identification of traits that confer drought tolerance is challenging because of dynamic risk-benefit ratios of varying cropping scenarios and the limited range of environments considered in most studies (Tardieu, 2011). Furthermore, the focus on a specific drought-tolerance traits and method of detection often excludes or disregards pleiotropic regulation of other traits. The encompassing trait to study drought-tolerance is typically connected to transpiration, which is the process of water uptake from the roots to the shoots and evaporation at the plants aerial components. For example, limited transpiration is a physiological consequence of a number of drought-tolerance traits such as stomatal closure in response to increasing vapor pressure deficit (VPD), inhibition of root water uptake in response to drying soil, and reduction of leaf area. Biophysically, these traits are connected through transpiration, whereby water uptake by roots is regulated by the water potential difference and hydraulic conductivity between the soil-root and leaf-air difference (Lobet *et al.*, 2014). Genetically, these traits most likely share pleiotropic regulatory pathways, such as abscisic acid signaling (McAdam and Brodribb, 2015). Chapter 4 describes the construction of a foundational framework for bioenergy sorghum crop modeling by parameterizing a sorghum crop model using field data for an energy sorghum hybrid; this model is able to take as input environmental parameters, like temperature and rainfall, and predict growth trajectories. We further extend the model to evaluate limiting transpiration traits, characterize benefits in water-limited environments in order to direct further analysis to determine ideotypes of sorghum

in target environments.

The work in this dissertation focuses on mathematically formalizing biological phenomena and their implementations to make predictions that can direct and accelerate discovery for sorghum crop improvement. The research herein illustrates this method to and its application to the field of agricultural genomics: (i) modeling genetic expectations to improve the association of genotype to phenotype, and (ii) modeling the physical consequences of phenotypes in their target environments and in turn its effect on crop productivity. The following three chapters reproduced manuscripts describing this approach and the insights obtained.

2. RESOLUTION OF GENETIC MAP EXPANSION CAUSED BY EXCESS HETEROZYGOSITY IN PLANT RECOMBINANT INBRED POPULATIONS [1]

2.1 Overview

Recombinant inbred populations of many plant species exhibit more heterozygosity than expected under the Mendelian model of segregation. This segregation distortion causes the overestimation of recombination frequencies and consequent genetic map expansion. Here we build upon existing genetic models of differential zygotic viability to model a heterozygote fitness term and calculate expected genotypic proportions in recombinant inbred populations propagated by selfing. We implement this model using the existing open-source genetic map construction code base for R/qtl to estimate recombination fractions. Lastly, we show that accounting for excess heterozygosity in a sorghum recombinant inbred mapping population shrinks the genetic map by 213 cM (a 13% decrease corresponding to 4.26 fewer recombinations per meiosis). More accurate estimates of linkage benefit linkage-based analyses used in the identification and utilization of causal genetic variation.

2.2 Introduction

Linkage maps, or genetic maps, are the relative ordering of and distance between genetic loci in terms of the frequency of recombination between them. Knowledge of the linkage between loci is useful for the identification and utilization of causal

[1]Reprinted from S. K. Truong, McCormick, R. F., Morishige, D.M., and J. E. Mullet, 2014. Resolution of genetic map expansion caused by excess heterozygosity in plant recombinant inbred populations. G3: Genes — Genomes — Genetics 4: 1963-1969 under the Creative Commons Attribution Unported License (`http://creativecommons.org/licenses/by/3.0`), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Copyright ©2014 Truong & McCormick *et al.*

genetic variation using techniques like map-based cloning, marker-assisted selection, and quantitative trait locus (QTL) mapping. Using this linkage information, the genotype of an observable locus (i.e. a marker) can be used to predict the genotype at proximal loci (e.g. a QTL), and the correct relative ordering of markers can be inferred for applications like genome assembly. As such, accurately calculating the linkage between markers is of practical importance.

For some plant species, including maize and pea, discrepancies in recombination frequencies exist between genetic maps calculated using recombination frequency estimates between markers and cytological maps calculated by observing cytological manifestations of recombination events with microscopy (Hall *et al.*, 1997a,b; Anderson *et al.*, 2003). In general, the genetic maps predict more recombination events per meiosis than the cytological maps observe, and cytological maps are considered to more accurately represent true recombination rates (King *et al.*, 2002). Two of the major factors contributing to this disparity include tight double recombination events and segregation distortion found in marker data (Sybenga, 1996; Knox and Ellis, 2002). Tight double recombinations are observed when an allele is found in a phase opposite to that of alleles from adjacent markers within a relatively short genetic distance (e.g. < 5 cM). The source of tight double recombinations is still an open question; they could arise from biological phenomena such as mutations or gene conversions, or they could be (and experiments have shown that they are generally are) genotyping errors (Broman and Weber, 2000; Broman *et al.*, 1998; Lincoln and Lander, 1992; Dib *et al.*, 1996). However, the frequency with which they are observed, even with disparate genotyping technologies, suggest that there may be an underlying biological process responsible for some of these tight double recombinations (Sybenga, 1996; Broman *et al.*, 1998). Regardless, because of the assumptions implicit in genetic map construction, tight double recombinations in

marker data greatly expand genetic maps. For the purposes of this paper, due to the dramatic map expansion caused by tight double recombinations and our current inability to conclusively identify their origins, we treat tight double recombinations as genotyping errors and set them to missing as commonly practiced in the literature (Lincoln and Lander, 1992; Broman *et al.*, 1998; Dib *et al.*, 1996).

Similarly, segregation distortion is a commonly observed phenomenon and can also affect the estimation of recombination frequencies. Segregation distortion is the observed deviation of a locus from the expected segregation ratio under the model of Mendelian inheritance, and it generally occurs as a consequence of unequal gametic or zygotic fitness (e.g. artificial selection, meiotic drive, etc.), or as a consequence of an error prone marker. While one solution for distorted markers is their removal, the removal of markers reduces genome coverage, and techniques have been developed to account for distorted markers by (i) integrating repeated observations in multiple populations (Wang *et al.*, 2005; Cloutier *et al.*, 2012) or (ii) modeling the differential viability of gametes or zygotes (Zhu *et al.*, 2007; Xu, 2008; Wu *et al.*, 2007; Lorieux *et al.*, 1995a,b). Additionally, information on segregation distortion can be used to aid in the identification of selection and QTL mapping (Xu, 2008).

Multiple reports have documented extensive segregation distortion in plant recombinant inbred populations manifesting mostly as excess heterozygosity, and occasionally as reduced heterozygosity (Table S1) (Knox and Ellis, 2002). While the source of distortion is not conclusively known, it is hypothesized to be the result of a general selective advantage (or disadvantage) of heterozygote genotypes. Despite the prevalence of heterozygosity in plant recombinant inbred lines (RILs), the techniques developed to incorporate distorted markers are not commonly employed. In the case of retaining markers based on multiple observations, this technique necessitates multiple RIL populations (Cloutier *et al.*, 2012) which may be too high a barrier given

some plant generation intervals. Existing methods to model the viability of each genotype differentially treats each marker pair, and so may suffer from the overfitting of large data sets without specific biological models; these have also only been shown for BC_1 and F_2 populations (Lorieux *et al.*, 1995a,b; Zhu *et al.*, 2007). In general, plant geneticists have constructed genetic maps of F_t populations, where t is the generation interval, by fitting observations to the expected genotype frequencies of a Mendelian fixed RIL model which relies on assumptions of complete fixation, no selection, and no mutation; this model is unable to account for proportions of heterozygosity maintained per generation other than 0.5. If the recombinant inbred population is treated as though all loci are fixed (as $t \to \infty$), yet more heterozygosity was maintained per generation than expected by Mendelian segregation on the way to fixation, then the recombination frequencies will be artificially overestimated; not accounting for excess heterozygosity underestimates the number of informative meioses that can occur prior to fixation. Additionally, treating RIL populations that have not yet reached fixation as fixed RILs results in the loss of genotypic information and makes incorrect assumptions when calculating recombination fractions. Lastly, in cases where the distortion occurs across the entire genome, such as for the sorghum mapping population used here, removal of distorted markers under a Mendelian F_t model would remove the majority of typed markers, causing a dramatic loss of genetic information.

Here, we build off an existing model of differential zygotic viability to incorporate a heterozygosity maintenance term for plant recombinant inbred populations and find a new solution for the genotype probabilities used to calculate recombination frequencies. We incorporate this calculation for expected genotype frequencies to account for different proportions of heterozygosity maintained per generation (other than 0.5) using the open-source genetic map construction code base from R/qtl

(Broman *et al.*, 2003), and report its efficacy in a simulated RIL population and a sorghum mapping population.

This modeling allows more accurate generation of genetic maps and retention of more genetic information by accounting for the biological phenomenon of differential fitness of heterozygous loci. More accurate estimations of recombination fractions, and thus linkage, will improve the accuracy of methods that use linkage information to detect and utilize causal genetic variants.

2.3 Results

2.3.1 Excess heterozygosity generally causes overestimation of recombination frequencies

To demonstrate how excess heterozygosity expands genetic maps, we plotted the estimated recombination frequency, $\hat{r}$, given genotype counts expected under conditions of excess heterozygosity for different recombination frequencies, r, estimated using the Mendelian model and using the derived heterozygosity model (see Materials and Methods for the derivation and implementation). When the genotype counts for the two markers arise from an excess heterozygosity model for an F_7 RIL population, accounting for the excess heterozygosity when calculating $\hat{r}$ correctly estimates the recombination frequency, r, underlying the data (Figure 2.1). However, using the Mendelian model to estimate $\hat{r}$ results in an overestimation relative to the recombination frequency underlying the data. Overestimation of $\hat{r}$ decreases as linkage increases ($r \leq 0.3$), and even these small overestimations between many pairs of markers lead to map expansion proportional to the genetic distance of the region with excess heterozygosity.

To further demonstrate the effects of excess heterozygosity, we simulated an F_7 RIL population of 1000 individuals with a 200 cM linkage group covered by

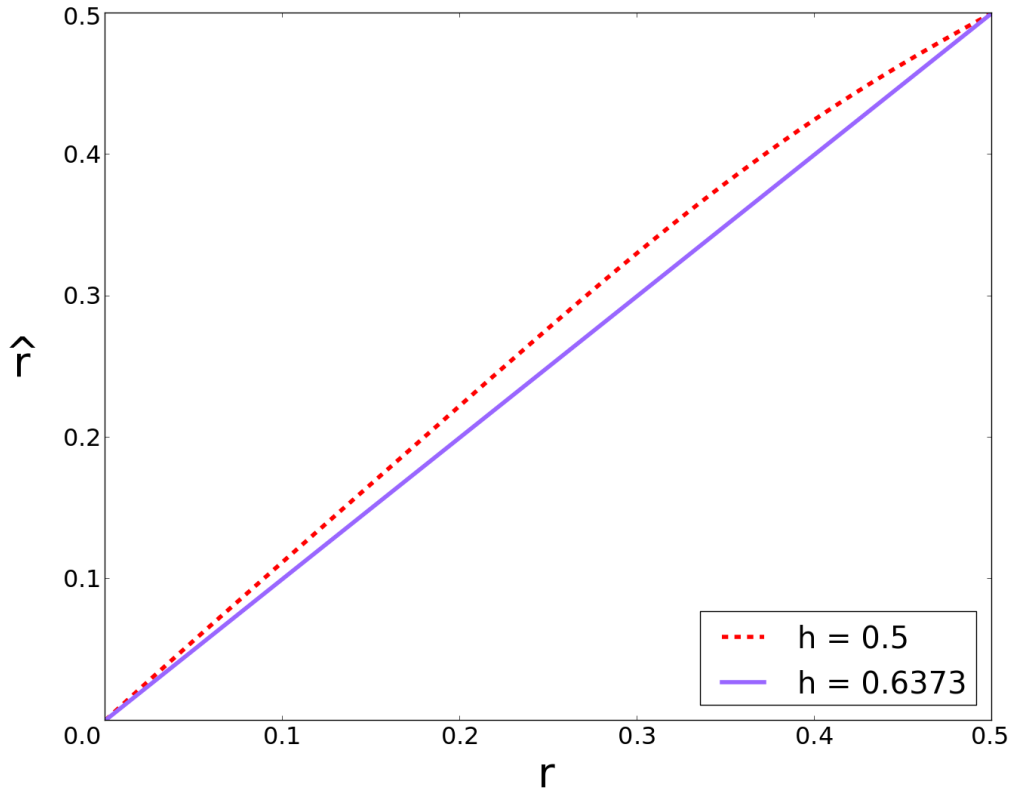

Figure 2.1: **Estimated recombination frequencies, $\hat{r}$, under excess heterozygosity and Mendelian models.** Recombination frequencies estimated from genotype frequencies under Mendelian expectations ($h = 0.5$) versus under modeling a global heterozygosity advantage ($h = 0.6373$) at generation $t = 7$ of a selfing population. This shows that if the population was retaining excess heterozygosity (at a rate of 63.73% each generation as opposed to the Mendelian 50%), then estimating recombination fractions under Mendelian expectations would lead to overestimation of the recombination frequency underlying the data and subsequent map expansion.

1000 markers under conditions of excess heterozygosity maintained per generation ($h = 0.6373$). Estimating recombination frequencies under a Mendelian model ($h = 0.5$) overestimates the map by 18.0% (236.0 cM), whereas accounting for excess heterozygosity in the genetic model yields a genetic map that differs from the

simulated distance by only 2.5% (204.9 cM) (Figure A.3).

2.3.2 Incorporation of a heterozygosity term into the genetic model shrinks a sorghum genetic map

To demonstrate that accounting for excess heterozygosity can shrink the genetic map of a plant recombinant inbred population (as postulated by Knox and Ellis in 2002 (Knox and Ellis, 2002)), we applied our method to a sorghum recombinant inbred population displaying excess heterozygosity (Burow *et al.*, 2011). The members of the population ranged from F_7 to F_9 and exhibit more than a 300% increase in heterozygosity relative to the expected heterozygosity given a Mendelian model: 6.7% observed after our quality control steps versus 1.6% given a Mendelian model for $t = 7$ (Figure 2.2). Heterozygosity was present at elevated levels throughout the genome relative to expectations under a Mendelian model, although some regions deviated notably from the average (Figures 2.3, A.1, and A.2.1). Previous reports estimating the genetic map as a RIL that has gone to fixation for this sorghum population range from 1,279 cM to 1,713 cM, a difference of 8.48 recombinations per meiosis (Table S2) (Hart *et al.*, 2001; Burow *et al.*, 2011; Peng *et al.*, 1999; Mace *et al.*, 2009; Menz *et al.*, 2002).

The genotype calls for this population were used to parameterize the heterozygosity term, h, by treating the population as an F_7 such that H_{F_7}=0.067 and h=0.6373 by equation 2.2; u and d were subsequently found by equation 2.3 (Materials and Methods). Figure 2.3 compares the genetic maps of the sorghum recombinant inbred population estimated as an F_7 under Mendelian expectations ($h = 0.5$) on the left and estimated under the excess heterozygosity model ($h = 0.6373$) on the right for each chromosome. Once excess heterozygosity is accounted for, the genetic map shrinks from 1,603.8 cM to 1,390.6 cM, a 213.2 cM difference corresponding to a 13%

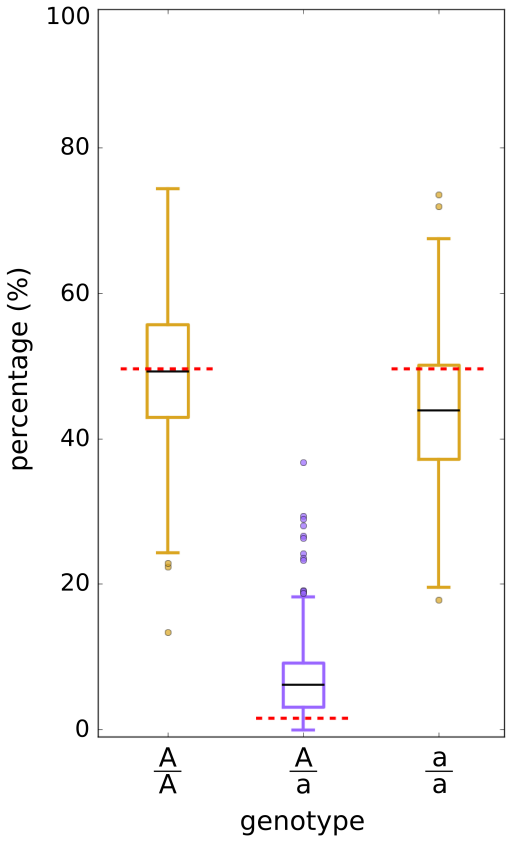

Figure 2.2: **Excess heterozygosity in a sorghum mapping population.** Box plot of genotype frequencies of 398 individuals of the BTx623×IS3620c recombinant inbred population. Each individual has a percentage of its genotypes that are homozygous or heterozygous for a BTx623 parental allele, A, and IS3620c parental allele, a. The red dashed lines represent the expected genotype frequencies under the assumptions of Mendelian segregation. The expected heterozygous frequency is lower than the median observed.

decrease, or 4.26 recombinations fewer recombinations per meiosis. As expected, the derived heterozygosity model behaves identically to the Mendelian model when $h = 0.5$ (Table 2.1).

2.4 Discussion

Observations that deviate from a model's expectations, such as segregation distortion caused by excess heterozygosity, will generally cause the model to generate inaccurate estimations; unsurprisingly, excess heterozygosity leads to unexpected map lengths when the genetic map is estimated under the assumptions of Mendelian segregation. Here we have shown that the excess heterozygosity present in a sorghum recombinant inbred population caused map expansion under Mendelian expectations. However there is no theoretical reason why excess heterozygosity could not also shrink the genetic map under certain conditions. If we had observed this recombinant inbred population in its F_3 stage and parametrized an $h = 0.6373$ (the same amount of heterozygosity maintained each generation), excess heterozygosity in this case would cause recombination frequencies to be underestimated under the Mendelian model (Figure A.1). While this does not agree with the idea that excess heterozygosity always causes map expansion (Knox and Ellis, 2002), this is not an unexpected result; our assumptions are dependent on a RIL approaching fixation ($t \to \infty$) in which case the longer maintenance of heterozygous loci provides more opportunities for recombination at the given loci. In other words, in the context of the genetic model derived in Materials and Methods, as $t \to \infty$, the proportion of class 2 genotypes ($\frac{\text{Ab}}{\text{Ab}}$ and $\frac{\text{aB}}{\text{aB}}$) will be larger for populations with excess heterozygosity than those following Mendelian expectations. Under our model and moderate values of h (e.g. $h = 0.6373$), excess heterozygosity is predicted to cause map shrinkage for small generation values (e.g. $t = 3$), and map expansion for larger generation values (e.g.

Chr	**est.rf()**	**est.rf.exHet**($h = 0.5$)	**est.rf.exHet**($h = 0.6373$)	**Burow et al. (2011)**
1	206.7	206.7	177.0	231.6
2	213.4	213.4	185.9	205.0
3	208.6	208.6	179.2	202.4
4	169.2	169.2	146.2	174.4
5	126.2	126.2	109.4	138.2
6	135.7	135.7	117.6	115.6
7	127.5	127.5	113.0	155.7
8	114.7	114.7	101.0	152.3
9	135.3	135.3	118.7	153.0
10	166.6	166.6	142.8	148.4
Total	1603.8	1603.8	1390.6	1676.6

Table 2.1: **Genetic maps estimated from the BTx623 x IS3620c sorghum recombinant inbred mapping population.** Except for the map reported by Burow et al. 2011 (which was treated as a fixed RIL), maps were estimated as a selfed F_7 population. The est.rf() function uses R/qtl's native recombination frequency calculations, whereas est.rf.exHet() uses the calculations detailed in the Materials and Methods with the respective h values. The map produced by Burow et al. in 2011 from a subset of the BTx623 x IS3620c population is provided as reference.

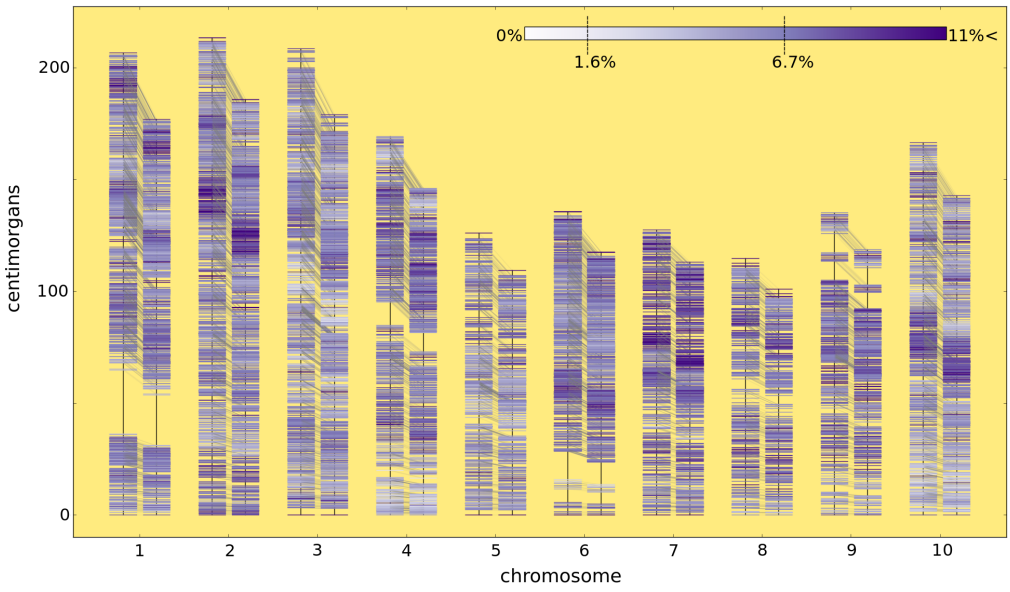

Figure 2.3: **Accounting for excess heterozygosity shrinks the sorghum genetic map.** This plot shows the genetic position of 10,081 markers for two genetic maps. For each chromosome, the genetic map on the left is calculated under the Mendelian segregation model. The genetic map on the right is calculated under the excess heterozygosity model. For all chromosomes (#1-10), the map shrinks by accounting for excess heterozygosity. The coloring of the markers correspond to the percentage of heterozygosity at that locus (no heterozygosity, white, to high ($> 11\%$) heterozygosity, purple). The expected heterozygosity of an F_7 RIL population is 1.6% and the observed heterozygosity in the BTx623×IS3620c population was 6.7% as depicted on the color bar. Faint gray lines connect a marker's position in one map with its corresponding position in the other map.

$t = 7$). The general case is simply that, when the observed genotype frequencies deviate from those predicted by the model, the estimated recombination fractions, $\hat{r}$, will be inaccurate.

Our modeling was done under the assumption that the amount of heterozygosity maintained by each generation is evenly distributed among markers. While our data for this population show that excess heterozygosity is present throughout the genome, there is also local variation (Figures 2.3 and A.1). This agrees with previous work showing that hybrid advantage and/or disadvantage can localize to specific loci in the genome (Li *et al.*, 1997), and in these cases it may be more appropriate to obtain an $h_{\text{marker pair}}$ from data H_{F_t} for each marker pair (which we derive in the Supplemental Information). However we chose not to implement this method for our mapping population to avoid overfitting the data and to maintain an expected global value with which the genetic map could be curated. We have additionally implemented an option to parameterize h for each linkage group based on the H of the genotypes for the linkage group, though this yielded little difference for our use case (Figure A.1). Ultimately, our solution strikes a balance between an *a priori* model based on Mendelian segregation and parameterizing the model based entirely on each marker pair. Future work may explore an intermediate approach based on estimating regional heterozygosity levels to paramaterize h for groups of markers; an examination of the biological mechanisms underlying these regional deviations from the global level of maintained heterozygosity is also merited (Figures A.1 and A.2.1).

We have made the implementation of this method available as a fork of the R/qtl code base at `https://github.com/MulletLab/qtl`, and provide examples for its use at `https://github.com/MulletLab/exHet_Supplement` so that it can be utilized where appropriate for conditions of excess heterozygosity; we are grateful to the R/qtl developers for making their code base accessible to be built upon (Broman

et al., 2003). We believe this work serves as an example of when to extend a genetic model to fit observations of biological phenomena that deviate from traditional expectations, and that the differential zygotic viability model (Zhu *et al.*, 2007) will serve as a useful base to modify as the mechanisms underlying segregation distortion become better understood. As genotyping technologies continue to improve, so too should the models we use to interpret the phenomena underlying the data. Deviations from traditional models, including segregation distortion and tight double recombinations, will need to be corrected to generate genetic maps that have reasonable agreement with the cytological maps calculated using microscopy to observe indicators of recombination events. More accurate genetic maps will improve linkage based analyses such as map-based cloning, marker-assisted selection, and QTL mapping, as well as assist marker ordering for genome assembly and provide better estimates of how recombination is distributed in the genome.

2.5 Materials and methods

2.5.1 Derivation and implementation of genetic model

The model for genetic map construction from genetic markers of a population with known pedigree and markers ordered on the basis of physical position with a reference genome is simplified into calculating recombination fractions between pairs of markers. Here we derive the quantitative genetic theory underlying the expected genotypes of a selfed population, F_t, given a proportion of heterozygosity retained that deviates from Mendelian segregation assumptions. To derive these equations we simultaneously extend and incorporate two models: (a) Bulmer's general solutions for genotype frequencies of self-fertilized populations based on the work of Haldane and Waddington at two linked loci (Haldane and Waddington, 1931; Bulmer *et al.*, 1980) and (b) a model for zygotic differential viability, where each genotype is assigned a

fitness (that may confer an advantage/disadvantage) in the F_2 progeny (Zhu *et al.*, 2007; Wu *et al.*, 2007).

2.5.1.1 Genotype frequencies of selfing populations

Before we model heterozygosity maintenance into estimating genotype frequencies, we will set up the familiar framework used to estimate genotype frequencies in a traditional recombinant inbred line. Consider two linked loci (or markers) α and β. Locus α has alleles A and a and locus β has alleles B and b. Suppose that the initial parental mating was $\frac{\mathrm{AB}}{\mathrm{AB}} \times \frac{\mathrm{ab}}{\mathrm{ab}}$, then in subsequent generations, F_t where $t \in \mathbb{N}$ is the generation interval, the family of individuals will contain a distribution of ten different genotypes, and for the initial condition of $t = 1$, all genotypes in the F_1 generation $\in \frac{\mathrm{AB}}{\mathrm{ab}}$ genotype. Furthermore, due to the symmetry of genotypes under self-fertilization, the genotype probabilities are reduced into p_{F_t} as five genotype classes (class i, $i \in [1, 2, 3, 4, 5]$) as described by Haldane and Waddington (1931), where $\forall\, t,\ \|p_{F_t}\| = 1$.

$$p_{F_t} = \begin{pmatrix} p(\text{class 1}) \\ p(\text{class 2}) \\ p(\text{class 3}) \\ p(\text{class 4}) \\ p(\text{class 5}) \end{pmatrix}_t = \begin{pmatrix} p(\text{genotypes}) \\ p(\frac{\mathrm{AB}}{\mathrm{AB}}) + p(\frac{\mathrm{ab}}{\mathrm{ab}}) \\ p(\frac{\mathrm{Ab}}{\mathrm{Ab}}) + p(\frac{\mathrm{aB}}{\mathrm{aB}}) \\ p(\frac{\mathrm{AB}}{\mathrm{aB}}) + p(\frac{\mathrm{Ab}}{\mathrm{ab}}) + p(\frac{\mathrm{AB}}{\mathrm{Ab}}) + p(\frac{\mathrm{aB}}{\mathrm{ab}}) \\ p(\frac{\mathrm{AB}}{\mathrm{ab}}) \\ p(\frac{\mathrm{Ab}}{\mathrm{aB}}) \end{pmatrix}_t$$

The transition from one class to another each generation is a Markov chain and is described through the transition probability matrix, $\mathbf{T}$, that takes into consideration the gametic outputs of each class for each meiosis event, such that for generation t

and the initial condition ${p_{F_1}}' = [0, 0, 0, 1, 0]$ we satisfy

$$ {p_{F_{t+1}}}' = \mathbf{T}{p_{F_t}}' \tag{2.1} $$

Thus far we've introduced the common modeling of genotype probabilities for the case of self-fertilized populations. Solving for the genotype probabilities is dependent on defining the transition probability matrix, $\mathbf{T}$. Under the assumption of Mendelian segregation, $\mathbf{T}$ is defined and so p_{F_t} has a general solution which has been implemented (Bulmer *et al.*, 1980; Lander *et al.*, 1987; Broman *et al.*, 2003). Furthermore, for the case of differential zygotic viability, $\mathbf{T}$ has been modeled and solved for an F_2 (Zhu *et al.*, 2007).

2.5.1.2 Modeling heterozygosity

Here we model a heterozygosity maintenance term for selfed recombinant inbred populations to account for viabilities of heterozygote genotypes that deviate from Mendelian segregation (where $h = 0.5$). To construct $\mathbf{T}$ that accounts for the proportion of heterozygosity maintained each generation, h, we will examine each class' expected transition from generation t to generation $t+1$ under a potential deviation of h from 0.5.

Transition from class 1 and class 2 is fixed: Class 1 and 2 are the ultimate absorption states as $t \to \infty$. For example, the probability of class $i \neq 1$ in generation $t+1$ given that the marker pair was in class 1 in generation t is zero. Once a marker pair is in either class 1 or class 2, it will remain there.

Transition from class 3 depends on h: Class 3 requires consideration of the segregation of only one marker that is heterozygous in generation t as the other marker will be homozygous and thus fixed in any subsequent generation after t.

Let H_{F_t} be the proportion of heterozygosity observed for all markers of an F_t

family and assume that the amount of heterozygosity maintained in all markers, h, is constant each generation. Then we can solve for h through the following relationship $h^{t-1} = H_{F_t}$. h will be modeled into the transition probability matrix as a modifier of expected segregation. For data H_{F_t},

$$[]h = e^{\frac{\ln(H_{F_t})}{t-1}}. \tag{2.2}$$

For the heterozygous marker α and h (in class 3) in generation F_t, the genotype probabilities for generation F_{t+1} will be dependent on the expected segregation of the alleles of marker α, class $1 : 2 : 3$, which is $\frac{1-h}{2} : \frac{1-h}{2} : h$.

Transition from class 4 and 5 depends on h and r: Class 4 and 5 requires consideration of the segregation of two markers that are heterozygous at generation t and the recombination frequency, r, between the two markers.

Similar to treatment of heterozygosity for one marker, we now apply the same heterozygosity term to both markers α and β. We model this within the context of zygotic differential viability, as shown in (Lorieux *et al.*, 1995b; Wu *et al.*, 2007). Assume for marker α that the viability of genotype Aa relative to AA or aa is u and the same u applies to the alleles of marker β. Then, the genotype probabilities in generation F_{t+1} from class 4, $\frac{\text{AB}}{\text{ab}}$, or class 5, $\frac{\text{Ab}}{\text{aB}}$, in generation F_t is dependent on the segregation of alleles of marker α and β, class $1 : 2 : 3 : 4 : 5$. From class 4 this ratio will be $\frac{2(1-r)^2}{d} : \frac{2r^2}{d} : \frac{8ur(1-r)}{d} : \frac{2u^2(1-r)^2}{d} : \frac{2u^2r^2}{d}$, and from class 5 this ratio will be $\frac{2r^2}{d} : \frac{2(1-r)^2}{d} : \frac{8ur(1-r)}{d} : \frac{2u^2r^2}{d} : \frac{2u^2(1-r)^2}{d}$, where $d = 2(1-r)^2 + 8ur(1-r) + 2r^2 + 2u^2[(1-r)^2 + r^2]$.

To model the amount of heterozygosity retained in generation t for a marker pair of class j, for $j \in [4, 5]$, in the previous generation $t-1$ we model h, calculated by

equation 2.2 as

$$[]h = \frac{1}{2}p(\text{class 3 }_{t+1} \,|\, \text{class } j \text{ }_{t}) + p(\text{class 4 }_{t+1} \,|\, \text{class } j \text{ }_{t}) + p(\text{class 5 }_{t+1} \,|\, \text{class } j \text{ }_{t}) \tag{2.3}$$

such that we can calculate u with variable r and subsequently d .

Transition probability matrix, $\boldsymbol{T}$*:* Incorporating the transition from a given class to all classes in every generation, we now have a transition probability matrix,

$$[]\mathbf{T} = \begin{array}{c} \\ \text{class 1 }_{t+1} \\ \text{class 2 }_{t+1} \\ \text{class 3 }_{t+1} \\ \text{class 4 }_{t+1} \\ \text{class 5 }_{t+1} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{class 1 }_{t} & \text{class 2 }_{t} & \text{class 3 }_{t} & \text{class 4 }_{t} & \text{class 5 }_{t} \end{array} \\ \begin{pmatrix} 1 & 0 & \frac{1-h}{2} & \frac{2(1-r)^2}{d} & \frac{2r^2}{d} \\ 0 & 1 & \frac{1-h}{2} & \frac{2r^2}{d} & \frac{2(1-r)^2}{d} \\ 0 & 0 & h & \frac{8ur(1-r)}{d} & \frac{8ur(1-r)}{d} \\ 0 & 0 & 0 & \frac{2u^2(1-r)^2}{d} & \frac{2u^2r^2}{d} \\ 0 & 0 & 0 & \frac{2u^2r^2}{d} & \frac{2u^2(1-r)^2}{d} \end{pmatrix} \end{array} \tag{2.4}$$

With $\mathbf{T}$ defined, we solve for the general solution of p_{F_t} by equation 2.1 and initial condition $p_{F_1}{}' = [0, 0, 0, 1, 0]$ and use p_{F_t} to fit recombination fractions (see Supplemental Materials for calculations and solution). When the expectations of segregation are in fact Mendelian, $h = 0.5$, then as expected the solution for genotype frequencies will reduce to the same ones solved for by Haldane and Waddington (1931).

2.5.1.3 Genotype frequencies with heterozygosity model

Given the theory derived for $p_{F_t}{}' = \mathbf{T}p_{F_{t-1}}{}'$ that is altered with a heterozygosity model, we solved for the general solution of p_{F_t}, genotype frequencies, using MATLAB (2010), and an M-file is provided in the Supplemental Materials to document all variables defined and calculations.

2.5.1.4 Implementation and simulation

Calculations of the genotype frequencies for proportions of heterozygosity maintained, h, other than 0.5 were implemented in C within a fork of the R/qtl v1.28.19 code base (Broman *et al.*, 2003). Specifically we used the golden section search algorithm as implemented in the R/qtl BC_sF_t tools (Shannon *et al.*, 2013) to estimate recombination fractions given genotype data for a marker pair. Map distances were calculated using the Haldane mapping function given the recombination fractions estimated from the golden section search.

The source code is available on GitHub as a forked R/qtl repository at `https://github.com/MulletLab/qtl`. The hetexp branch contains the new functions, including est.rf.exHet() that can be called from R similar to the existing est.rf() but with a heterozygosity term, h, passed to it. The est.rf.exHet() function can also estimate h on the basis of H for each linkage group. Example usage can be found at `https://github.com/MulletLab/exHet_supplemental`.

Genotypes for a 200 cM linkage group genotyped for 1000 individuals at 1000 markers were simulated under the derived heterozygosity model both a) without errors or missing data, and b) with 1% errors and 5% missing data. The code used to generate the datasets, the simulated datasets, and their respective results can be found at `https://github.com/MulletLab/exHet_supplemental`.

2.5.2 Plant materials and genotyping

The sorghum recombinant inbred mapping population, BTx623×IS3620c, were made available by the USDA-ARS Plant Genetic Resource and Conservation Unit, Griffin, GA.(Burow *et al.*, 2011). These F_{7-9} individuals were planted in College Station, TX fields in summer 2013. DNA was extracted from leaf tissue of 10-12 plants from seed stock of each RIL and prepared by Digital Genotyping (DG) with

restriction endonuclease *NgoMIV* (Morishige *et al.*, 2013). The DG templates were sequenced on Illumina HiSeq 2500 with 72 (or fewer) samples per lane.

Genotypes were generated from the sequenced reads of the recombinant inbred lines and their parents, BTx623 and IS3620c. The sequence reads were delivered already sorted on sample barcode, and they were checked for restriction sites using awk; where applicable, preprocessing was parallelized using GNU parallel (Tange, 2011). Reads were aligned to the sorghum reference genome (Sbi1) with BWA mem (v 0.7.5a) (Paterson *et al.*, 2009; Li and Durbin, 2010). Aligned reads were realigned around indels using the Genome Analysis Toolkit (GATK v3.1-1) and the Queue framework with IndelRealigner; individual GVCFs were generated using the HaplotypeCaller; and joint genotyping was performed using GenotypeGVCFs (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). Variants were hard filtered using VariantFiltration under the following criteria: DP $<$ 10; QD $<$ 5.0; MQ $<$ 30.0; MQRankSum $<$ -10.0; BaseQRankSum $<$ -10.0. The remaining variants were filtered to keep only biallelic variants for which the two parents, BTx623 and IS3620c, were each homozygous for different alleles and to keep only variants that were genotyped with a GQ score $\geq$ 20 in $\geq$ 25% of the samples. For these genotypes, the median depth of reads that passed the HaplotypeCaller's internal quality control metrics (i.e. the median sample-level DP annotation) was 17 reads. Genotypes with a GQ score $<$ 20 were set to missing, and those remaining were screened for tight double recombinations occurring within 2kbp; genotypes involved in a tight double recombination were set to missing. These variants and genotypes were used as the initial input for genetic map construction in R/qtl.

2.5.3 Genetic map construction

Genetic map construction was performed as an iterative process in R/qtl, starting with 424 individuals (RILs) genotyped at 12,836 SNP and indel markers. 2 individuals and 1,340 markers were removed due to high missingness levels ($\geq$ 60%). 7 individuals were removed due to sharing $\geq$ 90% of genotypes with another individuals. 703 markers were removed for being uninformative due to close proximity. 17 individuals were removed for having genotypic proportions far outside the distribution of most of the population members. The remaining 398 individuals and 10,793 markers had an overall 7.4% heterozygous genotypes. 7.4% was used as an initial H_0 to test for segregation distortion. Markers that deviated largely from a $463 : 74 : 463$ ratio (homozygous parent 1 : heterozygous : homozygous parent 2) by a chi-squared test ($p < 1 \times 10^{-15}$) were excluded; due to extreme segregation distortion for one parental allele on chromosome 1 caused by artificial selection of a known flowering time QTL, we dramatically relaxed the distortion threshold ($p < 1 \times 10^{-30}$) for chromosome 1 although a large gap still remains due to failing to retain markers in the region of the most severe distortion (Yang *et al.*, 2014). Following filtration, 398 individuals and 10,090 markers remained with 7.5% heterozygous genotypes. 7.5% was considered to better represent the true percentage of heterozygous genotypes in the dataset, and so we applied the segregation distortion test for the 10,793 markers with the updated H_1 of 7.5%; this retained 398 individuals typed at 10,091 markers. Of note, at our p-value thresholds, the expected Mendelian ratio ($H = 1.6\%$) retained only 4,512 markers, whereas the excess heterozygosity model ($H = 7.5\%$) retained 10,091 markers.

With the 398 individuals typed at 10,091 markers we then constructed an initial genetic map by estimating recombination fractions calculated under the excess het-

erozygosity model and R/qtl's implemented Haldane mapping function with markers grouped and ordered by their physical position on the Sbi1 reference genome. 10 markers on chromosome 6 were removed due to their incorrect placement on the Sbi1 reference assembly as indicated by inspection of recombination fractions and previous work (Morishige *et al.*, 2013). The genetic map was then re-estimated, and tight double recombinations less than or equal to 2.0 cM were removed. The proportion of heterozygosity at this point, $H = 6.7\%$, was used to estimate h for use in the final map estimation under the excess heterozygosity model; the same markers and genotypes were used for map estimation under the Mendelian model. Genetic maps were estimated directly from calculated pairwise recombination fractions for adjacent markers using R/qtl's implemented est.rf() and our implemented est.rf.exHet().

3. HARNESSING GENETIC VARIATION IN LEAF ANGLE TO INCREASE PRODUCTIVITY OF *SORGHUM BICOLOR*[1]

3.1 Overview

The efficiency with which a plant intercepts solar radiation is primarily determined by its architecture. Understanding the genetic regulation of plant architecture and how changes in architecture affect performance can be used to improve plant productivity. Leaf inclination angle, the angle at which a leaf emerges with respect to the stem, is a feature of plant architecture that influences how a plant canopy intercepts solar radiation. Here we identify extensive genetic variation for leaf inclination angle in the crop plant *Sorghum bicolor*, a C4 grass species used for production of grain, forage, and bioenergy. Multiple genetic loci that regulate leaf inclination angle were identified in recombinant inbred line populations of grain and bioenergy sorghum. Alleles of sorghum *dwarf-3*, a gene encoding a P-glycoprotein involved in polar auxin transport, are shown to change leaf inclination angle by up to 34 degrees (0.59 radians). The impact of heritable variation in leaf inclination angle on light interception in sorghum canopies was assessed using functional-structural plant models and field experiments. Smaller leaf inclination angles caused solar radiation to penetrate deeper into the canopy, and the resulting redistribution of light is predicted to increase the biomass yield potential of bioenergy sorghum by at least 3%. These results show that sorghum leaf angle is a heritable trait regulated by multiple loci, and that genetic variation in leaf angle can be used to modify plant architecture to improve sorghum crop performance.

[1]Reprinted with permission from "Harnessing genetic variation in leaf angle to increase productivity of *Sorghum bicolor*" by Truong, SK, McCormick, RF, Rooney, WL, and Mullet, JE, (2015) Genetics 201(3) 1229-1238, Copyright ©2015 by the Genetics Society of America.

3.2 Introduction

Sustainably increasing the productivity of crops on land currently used for agriculture without depleting natural resources is a global priority (Foley *et al.*, 2011; Drewry *et al.*, 2014). Improving the efficiency with which plants intercept solar radiation is one means to sustainably improve crop productivity. Leaf angle, or leaf erectness, is a plant canopy parameter that has drawn considerable attention due to the predicted improvement in photosynthetic efficiency and reduction in plant stress afforded by the redistribution of solar radiation from the top to lower levels of canopies (Tollenaar and Wu, 1999; Duvick, 2005a; Murchie and Reynolds, 2012; Drewry *et al.*, 2014; Mansfield and Mumm, 2014). Performance improvements predicted by theoretical models are corroborated by positive correlations between small leaf angles and cereal crop yields; post-green revolution rice cultivars have smaller leaf inclination angles and higher yields relative to their pre-green revolution predecessors (Yoshida, 1972; Sinclair and Sheehy, 1999; Sakamoto *et al.*, 2006), and modern maize is also characterized by small inclination angles as a consequence of selection for increased grain yield in breeding programs (Duvick, 2005a; Lee and Tollenaar, 2007; Hammer *et al.*, 2009; Tian *et al.*, 2011; Mansfield and Mumm, 2014).

Despite the association of leaf angle with increased productivity, its genetic basis remains to be fully characterized for many of the major grasses. In maize, *ligueless1* and *ligueless2* have been identified as major regulators of leaf angle that can improve plant productivity (Pendleton *et al.*, 1968; Lambert and Johnson, 1978; Moreno *et al.*, 1997; Walsh *et al.*, 1998). More than 40 additional quantitative trait loci (QTL) have been identified in the maize NAM population and RIL populations (McMullen *et al.*, 2009; Tian *et al.*, 2011; Li *et al.*, 2015). In rice, *osdwarf4-1* and *leaf inclination2* have been identified and shown to play roles in plant hormone responses that result

in changes in leaf angle (Sakamoto *et al.*, 2006; Zhao *et al.*, 2010). Progress in identifying leaf angle QTL has been made in sorghum, the 5th most widely produced grain and forage crop, however a gene that regulates leaf angle has yet to be identified as has been done in maize and rice (Hart *et al.*, 2001; Gill *et al.*, 2014; Perez *et al.*, 2014; Xin *et al.*, 2015).

The physiological basis for the impact of leaf inclination angles on yield may be explained by altered vertical distribution of solar radiation in the canopy. Leaf positioning, a factor influenced by leaf inclination angle, can maximize carbon gain by optimizing interception of photosynthetically active radiation (PAR) for canopy photosynthesis and by mitigating heat stress induced by excess infrared radiation (IR) (Zhu *et al.*, 2008; van Zanten *et al.*, 2010; Zhu *et al.*, 2010; Song *et al.*, 2013). In addition to mitigating the occurrence of excess radiation at the top of the canopy (Nobel, 2005; Long *et al.*, 2006; Zhu *et al.*, 2010; Mullet *et al.*, 2014), canopy architectures that have small upper leaf angles redistribute PAR more uniformly throughout the canopy, thereby reducing shade-induced senescence of lower leaves (Hurng *et al.*, 1986; Sinclair and Sheehy, 1999; van Zanten *et al.*, 2010; Song *et al.*, 2013). The resulting increase in green leaf area allows for greater accumulation of nitrogen, a possible rate-limiting factor during grain filling in modern high-yielding cultivars (Drouet and Bonhomme, 1999; Sinclair and Sheehy, 1999; Hammer *et al.*, 2009). More optimal vertical redistribution of solar radiation throughout the canopy can also allow for denser planting of grain crops, an important factor contributing to increased grain yield per hectare (Sinclair and Sheehy, 1999; Duvick, 2005b; Tian *et al.*, 2011; Mansfield and Mumm, 2014). For example, small leaf angles of the rice *osdwarf-4* mutant and the maize *liguleless2* mutant enabled higher density planting that increased biomass yield of the respective crops (Lambert and Johnson, 1978; Sakamoto *et al.*, 2006).

Given successes in other grass crops, there is strong motivation for identifying the genetic basis of leaf angle and determining its physiological consequences in sorghum, particularly for bioenergy sorghum. High biomass energy sorghum hybrids have long growing seasons and accumulate most of their biomass in tall (4 m) closed canopies (Rooney *et al.*, 2007; Olson *et al.*, 2012; Mullet *et al.*, 2014). Over the long bioenergy growing season, small daily improvements in energy conversion efficiency conferred by more optimal leaf angles could translate to large seasonal increases in biomass accumulation. We examined the genetic basis of leaf inclination angle in sorghum using both grain and bioenergy sorghum recombinant inbred line (RIL) populations and identified multiple QTL contributing to the regulation of leaf angle. Moreover, we demonstrate that a leaf angle QTL present in grain sorghum germplasm is caused by sorghum *dwarf-3* (a homologue of maize *br2*), and that the recessive allele of sorghum *dwarf-3* decreases leaf inclination angle up to 34 degrees (0.59 radians). Additionally, we use functional-structural plant modelling and field experiments to show that smaller leaf inclination angles cause solar radiation, including PAR, to penetrate deeper into energy sorghum canopies. The improvement in conversion efficiency afforded by the redistribution of PAR is predicted to increase the biomass yield of bioenergy sorghum over the growing season by at least 3%. Given these results, genetically optimizing leaf angle represents a promising way to sustainably increase sorghum productivity.

3.3 Methods

3.3.1 Genetic basis of leaf inclination angle

Two recombinant inbred line populations generated from biparental crosses of BTx623 x IS3620C (n=398) and R07018 x R07020 (n=96) were used to examine the genetic basis of leaf inclination angle in *Sorghum bicolor* (Burow *et al.*, 2011; Bartek

et al., 2012). Individuals from these populations were genotyped by sequencing using the restriction enzyme-based, reduced representation technique Digital Genotyping (Morishige *et al.*, 2013). Template DNA was prepared using the restriction enzyme NgoMIV for the BTx623 x IS3620C population and FseI for the R07018 x R07020 population. Libraries were sequenced with an Illumina Hi-Seq 2500. Reads were mapped to the sorghum reference sequence (Sbi1) with BWA (v0.7.5.a) (Paterson *et al.*, 2009; Li and Durbin, 2010). Aligned reads were processed with Picard (v1.108) and variant calls (SNPs and indels) were generated using the GATK (v3.2-2) by following the nave pipeline of the RIG workflow (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013; McCormick *et al.*, 2015). Subsequent genetic maps were generated using R/qtl (Broman *et al.*, 2003). The BTx623 x IS3620C genetic map contained 10,091 markers and was constructed as described in Truong *et al.* (2014), under a model of excess heterozygosity. Genotype data for R07018 x R07020 were quality controlled in a similar manner, and a genetic map containing 1,968 markers was estimated as an F5 using the BCsFt tools in R/qtl. Genetic maps and genotype data are available in the Supporting Information.

3.3.2 Phenotyping leaf inclination angle

The two experimental crosses used to study the genetic basis of leaf inclination angle, BTx623 x IS3620C and R07018 x R07020, were planted and phenotyped in College Station, TX in a greenhouse and experimental fields. Table 3.1 provides information on where, when, and the number of individuals from each population that were planted and phenotyped. Phenotypes used in heritability calculations and QTL mapping were the average of the biological replicates measured (Table 3.1). To phenotype leaf inclination angle, leaves were counted by starting from the youngest ligulated leaf to older leaves further from the top the plant. That is, at the time

of phenotyping, the youngest fully expanded leaf (below the whorl) was identified as leaf 1 and each subsequent leaf was numbered sequentially (e.g., leaf 3 was two leaves below leaf 1). Leaf inclination angle was measured with a protractor where the origin (pinpoint) was placed where the leaf midrib would meet the stem. The angle measured was between the stem to the adaxial midrib of the leaf (the axil).

3.3.3 Stepwise multiple-QTL mapping and heritability estimates

Stepwise multiple-QTL analysis was performed on the genotype and phenotype data described above to identify the quantitative trait loci (QTL) and epistatic interactions between them that best describe the genetic basis of variation in leaf inclination angle phenotypes. Phenotype data were normalized by Empirical Normal Quantile Transformation (Peng *et al.*, 2007). A single QTL analysis using the EM algorithm initially identified primary additive QTL, and this was used to seed model selection for multiple-QTL analysis. We employed the method of (Manichaikul *et al.*, 2009) for model selection as implemented in R/qtl (Broman *et al.*, 2003). This method considers the LOD score for a multiple-QTL model penalized by the complexity of the model, using penalty scores specific to each phenotype. We used computational resources on the WSGI cluster at Texas A&M University to calculate the penalties for main effects, heavy interactions, and light interactions; these penalties were calculated from 24,000 permutations of each phenotype from each population with genetic markers that were 2 cM apart (600 markers) to find a significance level of 5% in the context of a two-dimensional, two-QTL genome scan (penalized LOD scores are in Supporting Information).

We chose a parsimonious QTL model for each phenotype to report here (detailed statistics of each QTL model are in Supporting Information). The models chosen are less prone to false positives in our approach to QTL model selection as we required

additional QTL and interactions to increase the pLOD of the QTL model by at least its LOD penalties for main and light interactions, respectively. We did not choose an exhaustive QTL model, and less stringent QTL models for the phenotypes are possible.

In addition to variance explained by the QTL, we also estimated the variance explained by SNPs and indels across the genome for each phenotype using GCTA (Yang *et al.*, 2011). The set of genetic variants that were used as input to genetic map construction (prior to the quality control involved in map construction) were quality controlled using PLINK v1.90b3u (Purcell and Chang, 2014), and heritability was estimated using a genomic relationship matrix (GRM) and restricted maximum likelihood (REML) as implemented in GCTA. This analysis estimates the proportion of observed phenotypic variance-covariance relationships among the lines that can be explained by the whole-genome relatedness of the lines estimated by markers, providing an estimate of narrow-sense heritability (Yang *et al.*, 2013).

3.3.4 dw3 gene identification

The *dwarf-3* (*dw3*) gene is well known for its effect on stalk height in *Sorghum bicolor*. The common, non-functional *dw3* allele contains an 882 bp tandem duplication in exon 5 that causes the loss of function; the null *dw3* allele is genetically unstable and can revert to the functional *Dw3* allele when the 882 bp duplication is removed by unequal crossing over (Multani *et al.*, 2003). The reversion event is visually identifiable when it occurs in field plots of isogenic sorghum because the revertant is taller than its surrounding siblings. Seven *dw3* revertants were identified in rows of recombinant inbred lines of BTx623 (*dw3dw3*) x IS3620C (*Dw3Dw3*) planted in College Station, TX fields in 2014 (Figure B.1). The revertants and three of their non-revertant siblings were genotyped by polymerase chain reaction (PCR)

to amplify a region of *dw3*; genotypes that contained one or two copies of 882 bp DNA correspond to the functional and non-functional alleles, respectively. Primers used were designed by Barrero Farfan *et al.* (2012).

3.3.5 Calculation of the light extinction coefficient, k

Light interception throughout crop canopies is often formalized as an extinction coefficient, k (as derived in Beer-Lambert's Law), that relates the attenuation of light with properties of the material through which the light travels (Monteith and Moss, 1977; Monsi and Saeki, 2005; Nobel, 2005; Long *et al.*, 2006). Here, we use the equation

$$I(x) = I_0 \exp -kx \tag{3.1}$$

. Given data, $x_n, I(x_n)_{n=1}^{N}$, where $I(x)$ is the intensity (power per unit area) of radiation from the sunlight at depth x down the canopy, we can estimate light extinction coefficient, k, by fitting with the Levenberg-Marquardt algorithm for non-linear least-squares (Jones, 2001). A detailed explanation of the function and its adoption to describe the vertical distribution of sunlight in a crop canopy is provided in the Extended Materials and Methods. We use k as a descriptor to characterize and compare the distribution of light in simulated and field grown sorghum canopies with differing leaf inclination angles.

3.3.6 Virtual sorghum canopies

We constructed functional-structural plant models of sorghum and collected depth and incident light data, $x_n, I(x_n)_{n=1}^{N}$, in simulated light environments and calculated values of theoretical ks. The 3-dimensional virtual sorghum plants were constructed using Lindenmeyer systems in L-py (Boudon *et al.*, 2012). Lindenmeyer-systems provide a set of production rules whereby plant structural models are produced by

recursion through phytomers (Prusinkiewicz *et al.*, 2012). The virtual canopies were then illuminated by the nested radiosity model as implemented in CARIBU (Chelle and Andrieu, 1998; Chelle *et al.*, 2004; Pradal *et al.*, 2008) given light input that reflected solar conditions in College Station, TX (data retrieved from The United States Naval Observatory). Details on methods of construction and the scripts used can be found in the Supporting Information.

3.3.7 Field experimental sorghum canopies

Two sorghum RILs, RIL 63 and RIL 73, were identified as lines with large phenotypic variation in leaf angle when phenotyping the R07018 x R07020 RIL population grown under greenhouse conditions (Bartek *et al.*, 2012). To study the effect of leaf angle on light penetration at different depths in the plant canopy, two adjacent plots, one plot of RIL 63 and one plot of RIL 73, were planted in College Station, TX fields (W 96 20, N 30 37); each plot had 4 rows with row spacing of 0.76 m^2 and was planted at a density of 13.2 plants/m^2. To estimate k values for each of the two plots, light penetration at multiple layers in the canopy was measured for seven collections using two lifts and two PAR meters (LP-80, Decagon Devices, Inc., Pullman WA) to simultaneously take readings at the same canopy height in the two canopies (Extended Materials and Methods and Figure B.2 and Table B.1). All plot and plant measurements along with raw depth from the experimental field data, subsequent scaling employed, and their fits to Equation 3.1 are available in the Supporting Information.

3.3.8 Estimating potential biomass gain

The calculation of biomass yield gain is based on the gain of solar conversion efficiency on a per leaf basis of C4 plants. We used energy absorbed at each sequential leaf through the canopy (predicted by CARIBU model) and estimated the conversion

efficiency provided by Zhu *et al.* (2010). Conversion efficiency is modeled as a linear parameter in the Monteith equation for biomass (Zhu *et al.*, 2010), and thus the biomass yield gain is the same amount as the efficiency gain. The calculations made are available in the Supporting Information as a spreadsheet.

3.4 Results

3.4.1 Genetic regulation of sorghum leaf angle

Given that leaf inclination angle is associated with productivity increases in other crop plants (Sinclair and Sheehy, 1999; Duvick, 2005a; Tian *et al.*, 2011; Mansfield and Mumm, 2014), we sought to identify genetic loci that regulate leaf angle in sorghum. Identification of genetic loci that modulate leaf angle will enable both experimental analyses of leaf angle's effects on radiation use efficiency (RUE) and nitrogen status as well as deployment of favorable alleles into breeding programs. Genetic analyses were carried out using two RIL populations derived from the following biparental crosses: (a) grain sorghum lines BTx623 and IS3620C, and (b) late flowering bioenergy sorghum lines R07018 and R07020.

Leaf angle is developmentally regulated and changes based on leaf age; leaves at the top of the plant typically have small leaf angles and minimal variation that could be attributed to genetics (see Supporting Information). As such, the angles of the third and fourth (and fifth when not senesced) fully expanded ligulated leaves, counting from the uppermost fully expanded leaf at the top of the plant, were measured in the two RIL populations at varying stages of development grown in the field and in greenhouses (Table 3.1). Model selection of multiple-QTL mapping analyses as described in Manichaikul *et al.* (2009) identified 3 loci in the grain population (BTx623 x IS3620C) and 4 loci in the bioenergy population (R07018 x R07020) that affected leaf inclination angle (p ¡ 0.001; Table 2). In a given population, leaf number, and

environment, the proportion of phenotypic variance explained by the multiple-QTL model was always greater than 15%. Given that the small population sizes used here for mapping can inflate QTL variance estimates via the Beavis effect (Beavis, 1994, 1998; Xu, 2003), we also estimated the proportion of phenotypic variance explained by genome-wide SNPs and indels as described in Yang *et al.* (2011). Due to the relatedness of individuals in the populations, the proportion of phenotypic variance explained by the SNPs and indels using a genomic relationship matrix and restricted maximum likelihood (as implemented in GCTA) approximates the narrow-sense heritability of the trait (Yang *et al.*, 2013). Heritabilities of each leaf angle ranged from 9% to 62%, depending on the population, location, and plant age (Table 3.2). Heritabilities in the BTx623 x IS3620C population were much larger due to the large effect of the *dw3* locus discussed below; the dominant *Dw3* allele is fixed in R07018 and R07020 and does not segregate in the bioenergy population (Figure B.3).

Table 3.1: **Experimental crosses and phenotyping metadata.** This table contains the two biparental crosses used, the number of assayed, the location of plant growth, the number of biological replications phenotyped (i.e. number of plants), and the time at which measurements were acquired. Row plots in the field are planted mechanically and are 0.76 m apart and have planting density of approximately 13.2 plants/m^2.

cross generation	RILs (n)	location (College Station, TX)	planting		phenotyping	
			date	structure	date(s)	reps
BTx623 x IS3620c F_{6-8}	88	field	03/27/2013	1 row plot	06/20/2013, 06/26/2013, 06/30/2013	3
	336	greenhouse	05/27/2014	2 per pot	07/04/2014 - 07/05/2014	2
	7	field	04/08/2014	1 row plot	06/18/2014 - 06/19/2014	4
R07018 x R07020 F_5	94	greenhouse	07/30/2013	3 per pot	09/07/2013	3
	94	field	04/08/2014	1 row plot	06/27/2014	3
	94	field	04/08/2014	1 row plot	07/31/2014	3
	2	field	04/08/2014	4 row plots	07/22/2014 - 07/26/2014	4

3.4.2 dwarf-3 regulates leaf angle in sorghum.

Comparisons between QTL mapping results for leaf inclination angle and plant height in the BTx623 x IS3620C RIL population showed identical marker association with both traits across an interval of chromosome 7 corresponding to the *dwarf-3* (*dw3*) locus (data not shown). Previous reports also correlated alleles of *dw3* with stem dwarfing and leaf inclination angle (Hart *et al.*, 2001). Furthermore, *dw3* is known to segregate in this mapping population, where BTx623 (*dw3dw3*) carries the non-functional *dw3*, an unstable allele that contains an 882 bp direct duplication that can revert to *Dw3* by unequal crossing-over (Multani *et al.*, 2003), and IS3620C (*Dw3Dw3*) contains the functional *Dw3* allele. The large phenotypic effect of the dominant *Dw3* allele on height enables identification of tall revertants among dw3*dw3* RILs. Therefore, we screened a field-grown subset of the BTx623 x IS3620C RIL population and identified revertant plants in seven different RIL plots by their increased heights. We found that the revertant plants with increased height also displayed large leaf inclination angle relative to non-revertant plants (Figure 3.1 and Figure B.1). Genotypes at the *dw3* gene confirmed that the dominant *Dw3* allele was present in the tall revertant plants. *Dw3* revertant plants showed increases in leaf inclination angle of up to 34 degrees (0.59 radians) relative to *dw3* plants that were otherwise genetically identical individuals (Figure B.1). Thus, sorghum *dw3* has a pleiotropic effect on both height and leaf angle.

3.4.3 Leaf angle affects vertical light distribution in sorghum canopies.

Once genetic loci regulating leaf inclination angle were identified, we sought to determine if leaf inclination angle affects the distribution of solar radiation in the sorghum canopy and confers a functional difference in performance. To characterize and compare light interception by sorghum canopies with different leaf angles, the

Table 3.2: **Trait heritability and QTL of leaf inclination angle.** Estimates of trait heritability, h^2, were approximated from estimations of variance explained by SNPs and indels across the genome for each phenotype (Yang *et al.*, 2013). "***", "**", and "*" denotes p-values less than 0.0001, 0.05, and 0.1, respectively, for testing the null hypothesis that the trait is not heritable. All QTL models were chosen via model selection of multiple-QTL mapping analyses as described in Manichaikul *et al.* (2009). All QTL were additive and have p-values < 0.001. The biparental cross, phenotype location (field or greenhouse), number of days after planting (DAP), individuals (n), and leaf number (#) assayed in each experiment is given as metadata. The physical position (Sbi1, Paterson *et al.* (2009)) of largest LOD for each QTL and other statistics on the QTL models can be found in the Supporting Information. ˆ denotes LOD2 intervals that contain the *dw3* gene (chr7: 58.55-58.56 Mbp). "‡" denotes LOD2 intervals in which RIL 63 and RIL 73 (the RILs used in field measurements of light extinction) are not genotypically identical.

cross	phenotyped	leaf #	heritability (h^2)	QTL		
				Chr	LOD2 (Mbp)	variance (%)
BTx623 x IS3620c	fields; 90 DAP (n = 88)	3	0.62 ***	7	57.6 - 59.1 ˆ	37
		4	0.60 ***	1	55.6 - 60.4	12
				7	57.6 - 59.2 ˆ	38
	greenhouse; 38 DAP (n = 336)	3	0.54 ***	7	58.3 - 58.8 ˆ	37
		4	0.50 ***	5	2.7 - 4.8	5
				7	58.3 - 58.8 ˆ	33
R07018 x R07020	greenhouse; 39 DAP (n = 94)	3	0.39 ***	1	53.7 - 54.8 ‡	24
		4	0.31***	1	13.6 - 61.3 ‡	20
	fields; 80 DAP (n = 94)	3	0.12 **	7	55.2 - 58.3 ‡	16
		4	0.23 **	1	13.1 - 54.7 ‡	19
		5	0.33 **	7	13.1 - 15.2 ‡	20
				1	53.4 - 54.7 ‡	17
	fields; 114 DAP (n = 94)	3	0.09 *	3	5.6 - 61.1 ‡	16
		4	0.12 **	3	51.4 - 60.1	16
		5	0.14 **	3	56.5 - 59.5 ˆ	22

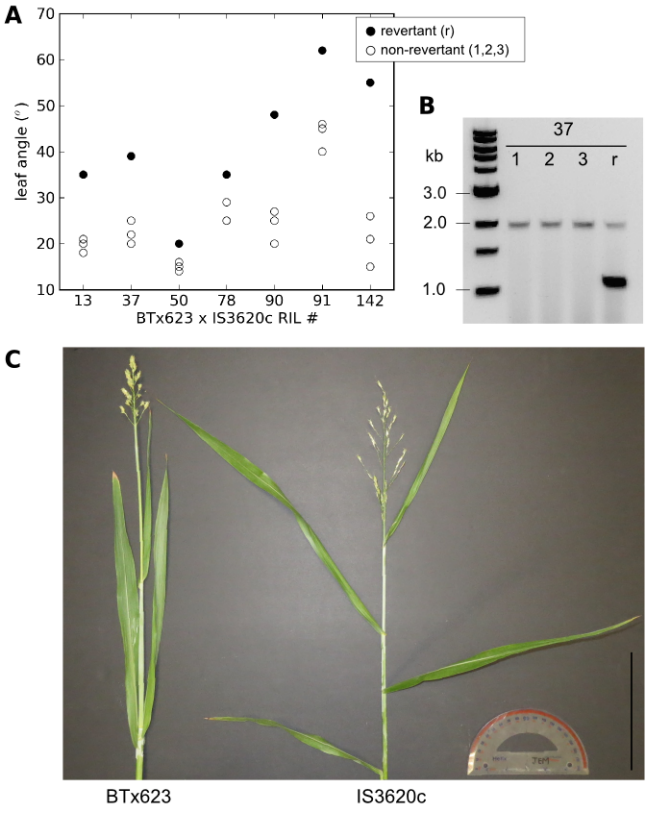

Figure 3.1: ***dw3* regulates leaf inclination angle. (A)** Leaf 3 inclination angles of revertants (r) in filled black circles and non-revertants (1, 2, 3) in empty white circles of seven RILs of BTx623 x IS3620C. **(B)** Genotypes of revertants and non-revertants of RIL 37 at the *dw3* locus using primers that flank the 882 bp tandem repeat that makes *dw3* non-functional; unequal crossing over can lead to excision of one repeat copy and spontaneous reversion to the functional *Dw3* allele (Multani *et al.*, 2003; Barrero Farfan *et al.*, 2012), see Figure B.1 for other RIL genotypes. **(C)** Parents, BTx623 and IS3620C, of the experimental cross in which revertants of *dw3* were identified grown under greenhouse conditions. The protractor used to measure all leaf inclination angles here is also shown, and the scale bar is 6 inches.

light extinction coefficients, k, of sorghum canopies were determined by modelling and using data from field measurements (Equation 3.1 in Methods). Given the same amount of available photosynthetically active radiation (PAR), decreases in k translate to less PAR intercepted by upper layers of the canopy and more PAR available for interception at lower levels of the canopy (i.e. closer to the ground) (Monsi and Saeki, 2005). We first compared the light extinction coefficient, k, of two simulated vegetative phase grain sorghum canopies that intercepted all available light (Chelle *et al.*, 2004; Pradal *et al.*, 2008; Boudon *et al.*, 2012). Plants in the simulated plots varying only in the leaf inclination angles of their fully expanded leaves were modelled under solar radiation conditions approximating a cloudless day in College Station, TX, on July 13, 2013 (retrieved from United States Naval Observatory, `aa.usno.navy.mil`) (Figure 3.2A). The canopies with larger leaf inclination angles had a k that was 2 times larger than canopies with smaller leaf inclination angles, indicating that light was distributed more uniformly and deeper into the canopies of plants with small leaf inclination angles (Figure 3.2B).

To determine the extent that leaf inclination angle alters light distribution in canopies under field conditions, we identified two recombinant inbred lines (RILs) from a biparental cross of bioenergy sorghum lines (R07018 x R07020), RIL 63 and RIL 73, that were phenotypically similar in all leaf traits measured except for leaf inclination angle (Figure B.4). RIL 63 had smaller leaf inclination angle at all phytomers compared to RIL 73 (Figure B.4). Notably, RIL 63 and RIL 73 from the R07018 x R07020 population had genotypic differences in 3 of the 4 QTL intervals affecting leaf angle identified by multiple-QTL mapping (indicated by ‡ in Table 3.2); these genetic differences may underlie the large phenotypic differences in leaf inclination angle observed in these RILs. The two RILs were phenotypically similar with respect to stand density, leaf morphology, phytomer stem diameter and spac-

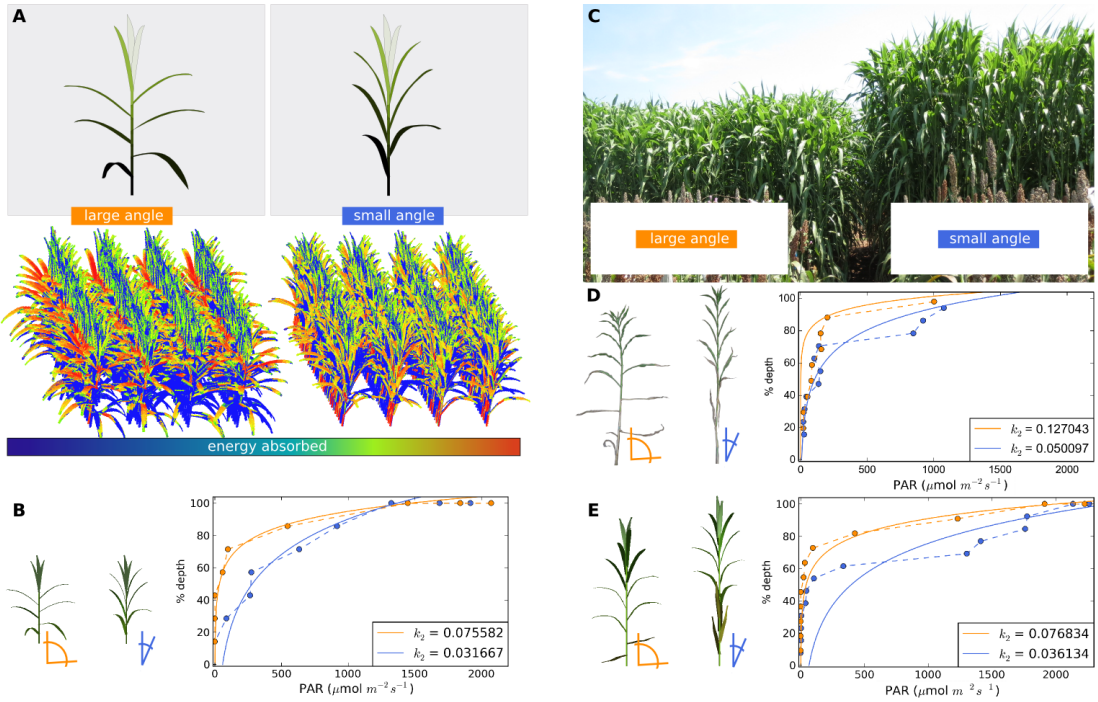

Figure 3.2: **Leaf inclination angle affects light distribution in sorghum canopies.** **(A)** Virtual sorghum plants and sorghum plots that vary in their leaf inclination angles (orange represents a larger leaf inclination angle relative to blue). **(B)** Light extinction curves for virtual plots from panel **A** simulated under solar conditions representing 13:00 on July 13, 2013 in College Station, TX (W 96 20, N 30 37). **(C)** Field plots of RIL 63 and RIL 73 from which light measurements were taken. **(D)** Light extinction curves for field plots from panel c assayed around 15:30 on July 22, 2014 in College Station, TX. **(E)** Light extinction curves for virtual plots representing RIL 63 and RIL 73 simulated under solar conditions representing 15:30 on July 22, 2014 in College Station, TX. In both simulation and field studies, plots with smaller leaf inclination angles fit a smaller light extinction coefficient, k, relative to plots with larger leaf inclination angles (panels **B**, **D**, and **E**). k values are derived from fits to Eq. 3.1 (see Methods).

ing, and biomass per unit height (Figure B.4). Notably, RIL 63 and RIL 73 varied with respect to the number of phytomers, leading to a difference in height and total biomass. Because of this height difference, measurements of k were scaled to percentage of canopy depth rather than absolute depth, although relative ordering of k values remains the same whether depth is scaled or not (Table 3.1). Four row plots of RIL 63 and RIL 73 were planted in the field at College Station, TX in 2014 and analyzed after canopy closure in July. PAR was measured throughout the canopy seven times over the course of four days using two PAR meters mounted on a platform that could be moved to varying heights in the canopies (Figure B.2 and Table B.1). Six of the data sets showed that the plots containing plants with smaller leaf inclination angle distributed PAR to greater depths in the canopy, with two of the seven showing that the k of the large leaf inclination angle plots is more than twice as large as the k of the small leaf inclination angle plot (Figure 3.2D and Table B.1).

While RIL 63 and RIL 73 were phenotypically similar for most measured traits and genotypically different at most detectable leaf angle QTL (Table 3.2 and Figure B.4), the possibility remained that their differences in light canopy penetration arose from factors not captured by the experiment such as additional QTL regulating plant architecture. To investigate this possibility, simulated plants were parameterized using the measured phenotypes of RIL 63 and RIL 73, including differences in phytomer number (and consequently, height); traits not measured, such as leaf curvature, were kept constant, and the simulated plants were illuminated using field conditions. The light distribution plots generated by simulations were qualitatively similar to the field results (Figure 3.2E and 3.2D, respectively). The possibility that differences in plant height and not leaf inclination angle were responsible for the smaller k of the small angle plant was ruled out by removing phytomers so that the plant with small leaf angles, simulated RIL 63, was the same height as the plant

with large leaf angles, simulated RIL 73; large leaf angles still caused k to be larger than small leaf angles (Figure B.5). Fitting of the light extinction function (Equation 3.1 in Methods) to the data revealed that the extinction model may not fully capture changes that occur in lower portions of the canopy (Figure 3.2D and 3.2E). This suggests that the light extinction models assumption of a homogenous canopy is an over-simplification (Song *et al.*, 2013), and factors like leaf orientation or leaf senescence may need to be considered in future adjustments to the model.

We next determined if the redistribution of incident PAR would be predicted to have a functional impact on biomass accumulation. Using estimates of conversion efficiency obtained from Zhu *et al.* (2010), we found that the redistribution of incident PAR in the simulated canopy with small leaf angles would cause an increase in overall conversion efficiency relative to the large leaf angle canopy. From this, we extrapolated a conservative estimate of biomass accumulation during a bioenergy sorghum growing season without water or nutrient limitations. For the virtual sorghum and lighting in Figure 3.2A, the conversion efficiency of the canopy with smaller leaf angles is predicted to be 1.0436 of the conversion efficiency of the canopy with large leaf angles. If we further assume that (i) the 4% gain in conversion efficiency is realized for 4 hours (midday) per 14 hour days, and (ii) the effect calculated is applicable to the duration of vegetative closed canopy growth, then given a bioenergy sorghum growing season where 140 days are in the vegetative closed canopy out of its 200 day growing season (Olson *et al.*, 2012), we predict an overall increase of 3% conversion efficiency over the entire growing season. Thus, under these conditions, the canopy with smaller leaf angles has the potential to accumulate 3% more biomass than the canopy with large leaf angles. Since these performance differences are predicted from morphologically reasonable canopy parameters, leaf angle represents a realistic target to improve biomass yields without increasing input.

3.5 Discussion

This study has shown that leaf inclination angles are genetically regulated in Sorghum bicolor, and that leaf angle alters the vertical distribution of solar radiation in closed canopies under field conditions. The improved distribution of solar radiation is predicted to improve photosynthetic conversion efficiency of the canopy, resulting in a 3% gain in biomass yield for bioenergy sorghum. We found that the phenotypic difference in leaf angle between RIL 63 and RIL 73 corresponded with genotypic differences at most leaf angle QTL identified in the population from which they originated, and that RIL 63 (with smaller leaf inclination angles) had a smaller light extinction coefficient (k) than RIL 73. Interestingly, RIL 63 also had more green leaf area and greater biomass than RIL 73. Additionally, while RIL 63 was taller, the length of each phytomer and biomass per unit height were similar (Figure B.4), indicating that RIL 63 had an increased rate of phytomer production. This increase in phytomers and, in turn, biomass, is consistent with the possibility that the small leaf angles of RIL 63 improved light interception and canopy conversion efficiency relative to RIL73. Leaf appearance rate, or phyllochron, varies among sorghum genotypes during development and in response to the environment (Clerget *et al.*, 2008; Van Oosterom *et al.*, 2011). In maize, field radiation intensity was negatively correlated with phyllochron (Birch *et al.*, 1998). Determining if the increased phyllochron in small leaf angle genotypes is a consequence of (i) differential genetic regulation of phyllochron, (ii) lower radiation interception by upper leaves, and/or (iii) increased photosynthetic conversion efficiency of the canopy will be investigated in future experiments. Additionally, prior work in maize and rice predict that the small leaf angles of RIL 63 will enable higher biomass accumulation at higher planting densities relative to RIL 73, so the effects of planting density should also be

investigated (Lambert and Johnson, 1978; Sakamoto *et al.*, 2006).

Identification of *dwarf-3*, which encodes a P-glycoprotein (PGP) that modulates polar auxin transport, as a regulator of leaf inclination angle by up to 34 degrees (0.59 radians) in sorghum was shown by analysis of naturally occurring revertants. As with height, the action of the null *dw3* allele on leaf angle is likely a consequence of reduced polar auxin transport from shoot apical meristem (Multani *et al.*, 2003; Knöller *et al.*, 2010). Auxin was shown to regulate the establishment and propagation of the preligule band in maize plants with null alleles of *liguleless1*, *liguleless2*, and *liguleless narrow* (Moon *et al.*, 2013), and PGP genes interact with PIN genes to influence local auxin distribution (Blakeslee *et al.*, 2007; Mravec *et al.*, 2008). Furthermore, rice *LAZY1*, a polar auxin transport repressor, regulates rice tiller angle via shoot gravitropism (Li *et al.*, 2007; Dong *et al.*, 2013), and maize *ZmCLA4*, an ortholog of rice *LAZY1*, increases shoot gravitropism and leaf angle at elevated expression levels (Zhang *et al.*, 2014). Additional work will be necessary to determine the mechanism by which *dw3* alleles regulate leaf angle. *dw3* also regulates sorghum height, and the pleiotropic effects of *dw3* explain observations from sorghum literature including a positive correlations between height and leaf angle in a sorghum grain association panel, and increased leaf angle in some *dw3* NILs (George-Jaeggli *et al.*, 2013; Perez *et al.*, 2014).

While *dw3* had the largest effect on leaf angle of the loci reported here, the utility of *dw3* for modulating leaf angle to improve sorghum productivity may be limited by its pleiotropic effects on height. Previous work has shown that the reduction in shoot biomass caused by *dw3* reduces radiation use efficiency (RUE) and grain yield, which would likely offset any benefit afforded by small leaf angle for grain and bioenergy applications (George-Jaeggli *et al.*, 2011, 2013). If the reduced productivity of grain sorghum with recessive *dw3* is due to a reduction in stem reserves as proposed by

George-Jaeggli *et al.* (2011), any conversion efficiency benefits conferred by leaf angle would be constrained by the sink limitation caused by reduced stem size. The efficiency benefits of small leaf angles caused by the null *dw3* allele may not be observed until recessive *dw3* is put into a genetic background that is not sink limited. For bioenergy applications, it is likely that the reduction in height and biomass caused by recessive *dw3* outweighs any productivity gains afforded by leaf angle, and other means of reducing leaf angle will be necessary. Fortunately, the genetic architecture underlying RIL 63 is promising since small leaf angles were associated with increased height and biomass.

In nature, large leaf angles may have been under positive selection because shading nearby plants would reduce competitors ability to compete for light and nutrients (Schmitt *et al.*, 2003; Drewry *et al.*, 2014). In monoculture grain cropping systems, small leaf angle has been under positive selection because this trait enables higher planting density and higher grain yield, presumably due to the benefits of improved radiation use efficiency and increased nitrogen content of canopies (Drewry *et al.*, 2014; Warnasooriya and Brutnell, 2014). The design of C4 energy grass crops is at an early phase of development and could benefit from improved radiation use efficiency afforded by small leaf angles (Mullet *et al.*, 2014). Energy sorghum and grain sorghum canopies close 60-75 days after seedling emergence, approximately when grain sorghum reaches anthesis. In contrast, energy sorghum remains vegetative following canopy closure for an additional 140 days and plants retain a whorl of developing leaves at the top of the canopy that have very small angles (Olson *et al.*, 2012). In grain sorghum, from anthesis through grain maturity, canopies lack the whorl of leaves with small angles. As such, we expect that the potential effects of leaf angles on yield will vary dependent on the duration of developmental phases of the crop.

The current study identified significant genetic variation for leaf angle in sorghum germplasm that can be deployed in energy, forage, and grain sorghum breeding programs. The simulation and field results support the conclusion that smaller leaf inclination angles cause a redistribution of solar radiation in closed canopies resulting in greater photosynthetic conversion efficiency and greater biomass yield potential. Large-scale studies of energy sorghum hybrids differing in leaf angles will be required to determine the extent to which this trait can increase biomass yield under field conditions.

4. BIOENERGY SORGHUM CROP MODEL PREDICTS VPD-LIMITED TRANSPIRATION TRAITS ENHANCE BIOMASS YIELD IN WATER-LIMITED ENVIRONMENTS[1]

4.1 Overview

Sorghum bicolor is a versatile, drought resilient crop used for production of grain, forage, sugar, and biomass. Bioenergy sorghum is targeted for production in drought-prone annual cropland not optimal for food production, and traits that improve plant water capture and water use are necessary to maximize productivity. A crop modeling framework, APSIM, used extensively for analysis of grain sorghum, was adapted to predict the growth and biomass yield of energy sorghum. APSIM simulations of energy sorghum development and biomass accumulation replicated results from field experiments across multiple years, patterns of rainfall, and irrigation schemes. Additionally, APSIM was extended to enable modeling of VPD-limited transpiration traits that limit crop water use under high vapor pressure deficits (VPDs). The response of transpiration rate to increasing VPD was modeled as a linear response until a VPD threshold was reached, at which the slope of the response changes, representing the range of responses to VPD observed in sorghum germplasm. Simulation results indicated that the VPD-limited transpiration trait is most beneficial in hot and dry regions of production where crops are exposed to extended periods without rainfall or terminal drought. In these settings, slower but more efficient transpiration increases biomass yield and prevents or delays the onset of plant dehydration and leaf senescence. The VPD-limited transpiration responses observed in

[1]Submitted manuscript "Bioenergy sorghum crop model predicts VPD-limited transpiration traits enhance biomass yield in water-limited environments" by Truong, SK, McCormick, RF, and Mullet, JE, (2016)

sorghum germplasm increased biomass accumulation by 20% in dry years, and the ability to drastically reduce transpiration under high VPD conditions could increase biomass by 6% on average across all years. This work indicates that the productivity and resilience of bioenergy sorghum grown in water-limited environments could be further enhanced by optimal deployment of VPD-limited transpiration traits. The energy sorghum model and VPD-limited transpiration trait implementation are made available to simulate performance in other target environments.

4.2 Introduction

*Sorghum bicolor*is an agronomically versatile grass crop grown for multiple production applications and has recently emerged as a promising lignocellulosic biomass crop. Due in part to its drought and heat tolerance, sorghum is capable of economically viable yield on land not optimal for production of many annual food crops (Borrell *et al.*, 2000a,b, 2014; Doggett, 1988; Kimber, 2000). Combined with C4 photosynthesis, high radiation interception and use efficiency, annual cropping seasons that permit rotations and adjustments for economic conditions, and the species amenability for genetic improvement of hybrids, these traits make energy sorghum a productive bioenergy crop for commercial production of biofuels and bio-products (Gill *et al.*, 2014; Mullet *et al.*, 2014; Rooney *et al.*, 2007). Constructing and breeding sorghum for bioenergy applications is a relatively new endeavor and many opportunities exist to improve plant productivity that have not been explored and evaluated. One approach to increase the rate of bioenergy sorghum crop improvement is to prioritize physiological traits for genetic engineering and breeding by using models to identify traits with significant yield benefits.

The Agricultural Production System SIMulator (APSIM) is an agricultural modeling framework that enables predictions of the growth and productivity of crop

species under varying environmental and management conditions (Holzworth *et al.*, 2014; Keating *et al.*, 2003; McCown *et al.*, 1996). The sorghum module in APSIM has benefitted from continuous development and improvement since 1994 (Hammer and Muchow, 1994), incorporating established models of sorghum phenology, canopy development, growth, and nitrogen use (Rosenthal *et al.*, 1987; Sinclair *et al.*, 1984; Birch *et al.*, 1990; Chapman *et al.*, 2000a,b; Hammer and Muchow, 1994). APSIM has also been enhanced to model complex adaptive traits and genotype to phenotype predictions (Hammer *et al.*, 2010). In addition to its record of providing accurate predictions of sorghum development (Hammer and Muchow, 1994; Kholova *et al.*, 2014; Kumar *et al.*, 2009; Lobell *et al.*, 2015), its modular design provides a flexible platform for examining the impact of variation in traits and physiological processes on plant development. To date, APSIM has been used predominately for modeling grain crops and additional work has demonstrated its efficacy in sugarcane (Nair *et al.*, 2012). As such, APSIM should be similarly effective for examining biomass accumulation in bioenergy sorghum with long growth durations. Therefore, we extended the grain sorghum model of APSIM to bioenergy sorghum and examined its performance relative to experimental data from field experiments across multiple years and patterns of rainfall and water availability. Additionally, APSIM was used to investigate the potential impact of limiting transpiration under high vapor pressure deficit (VPD) in bioenergy sorghum.

VPD describes the pressure difference between the leaf and its surrounding air, and limiting transpiration under high VPD has been predicted to contribute to the water use efficiency of grain sorghum and corn (Kholova *et al.*, 2014; Messina *et al.*, 2015; Sinclair *et al.*, 2005). The VPD-limited transpiration trait imposes a restriction on transpiration rate, enabling water conservation when VPD becomes sufficiently large (McAdam and Brodribb, 2015; Parent *et al.*, 2009). Previous experiments

have demonstrated that VPD-limited transpiration is a genetically regulated trait in sorghum, and some genotypes display a linear increase in transpiration rate with increasing VPD, whereas other genotypes display a VPD breakpoint, defined as the VPD at which the slope of the linear response between VPD and transpiration rate changes (Choudhary *et al.*, 2013; Choudhary and Sinclair, 2014; Gholipoor *et al.*, 2010; Riar *et al.*, 2015). These sorghum genotypes respond to high VPDs by reducing their transpiration rates, effectively limiting water loss under environmental conditions that result in low transpiration efficiencies, thereby increasing overall crop water use efficiency.

The VPD-limited transpiration trait may be especially beneficial for bioenergy sorghum because of the crops relatively long growing season (approx. 200 days). Any daily advantage that the trait confers would potentially be compounded over the long vegetative growth phases (Rooney *et al.*, 2007; Mullet *et al.*, 2014; Murphy *et al.*, 2011, 2014; Rooney and Aydin, 1999). Moreover, enhancing the ability of energy sorghum to tolerate periods of water limitation of varying length and onset during a long growing season will enable this crop to utilize intermittent rainfall to capture more water for growth and biomass accumulation. Since energy sorghum production is currently targeted for regions that are subject to intermittent water deficit and future climate change predicts that increases in VPD will be detrimental to vegetation (Lobell *et al.*, 2014; McDowell *et al.*, 2016), determining the potential impact of the VPD-limited transpiration trait on energy sorghum biomass yield and resilience is of great interest.

Previous approaches to incorporate a limited transpiration trait into crop models imposed a VPD-independent and a VPD-dependent maximum transpiration rate in grain sorghum and maize, respectively. Implementations of these models demonstrated that limited transpiration can result in improved grain yield in drought and

heat stressed environments by shifting water utilization from the vegetative and booting phases to the anthesis and post anthesis phases where yield is more sensitive to water deficit (Kholova *et al.*, 2014; Messina *et al.*, 2015; Sinclair *et al.*, 2005). In grain sorghum, VPD-limited transpiration was modeled as a maximum transpiration rate per unit leaf area, such that the transpiration rate would plateau at the designated maximum regardless of further increases in VPD (Kholova *et al.*, 2014; Sinclair *et al.*, 2005). In maize, limited transpiration was modeled as a piecewise function whereby, at or above a designated VPD (the VPD breakpoint), transpiration rate would plateau and not increase further (Messina *et al.*, 2015). These models are characteristic of some sorghum genotypes but do not capture the full range of VPD modulated transpiration responses reported for sorghum (Choudhary *et al.*, 2013; Choudhary and Sinclair, 2014; Gholipoor *et al.*, 2010; Riar *et al.*, 2015). To capture this additional complexity, we implemented a dynamic VPD-limited transpiration modification as part of an energy sorghum model in APSIM and used this modification to examine the effects of a range of VPD-limited transpiration traits on biomass accumulation. Adapting APSIM for bioenergy sorghum and VPD-limited transpiration enables predictions of energy sorghum phenology, biomass accumulation in a range of environments and agronomic practices, and analyses of the impact of the VPD-limited transpiration on biomass accumulation.

4.3 Methods

4.3.1 Crop model simulations for energy sorghum in APSIM

The daily progression of sorghum biomass accumulation given environmental data and management practices was simulated using Agricultural Production Systems Simulator (APSIM) (Holzworth *et al.*, 2014; Keating *et al.*, 2003; McCown *et al.*, 1996). The sorghum module distributed with APSIM 7.7 (www.apsim.info) incorpo-

rates the advances described by Hammer *et al.* (2010) and was used in this study. A generalized energy sorghum genotype was constructed based on the results of Olson *et al.* (2012) for the energy sorghum hybrid TX08001 by adjusting the parameters of a late maturity grain sorghum model described in APSIMs online documentation. Table 4.1 shows the parameters adjusted to parameterize the energy sorghum model, texas_energy. While the sorghum module in APSIM is capable of modeling photoperiod sensitivity (Hammer *et al.*, 2010), it is not designed to model the photoperiod sensitivity of lines that do not flower in long days regardless of accumulated thermal degree days. The implementation of photoperiod sensitivity in the APSIM sorghum module serves to extend thermal time in the vegetative phase and is therefore unable to replicate the physiological basis of flowering responses of sorghum genotypes like TX08001 that, once the minimum number of thermal degree days have accumulated, requires exposure to a sufficiently short photoperiod to initiate flowering (Holzworth *et al.*, 2014; Murphy *et al.*, 2011; Rooney *et al.*, 2007). Due to this limitation, the thermal time (e.g. photoperiod slope) parameters used were set to be consistent with observed time to floral initiation for TX08001 in College Station, Texas, but they will not generalize to other latitudes.

For environmental data, the daily maximum temperature, minimum temperature, and precipitation data for College Station, TX (Latitude 30.58917, Longitude -97.36472) from the beginning of year 2000 to the end of year 2014 were obtained from the Daily Global Historical Climatology Network, GHCN-DAILY, (Menne et al 2012a, 2012b, access date: January 2016). The maximum and minimum radiation per day were obtained from Maximum and Minimum Radiation Incident On An Equator-pointed Tilted Surface (kWh/m2/day) from NASA Surface meteorology and Solar Energy (SSE) data set for Latitude 30.601, Longitude 96.314 (`https://eosweb.larc.nasa.gov/`). The tav_amp APSIM function was used to

Table 4.1: **Parameter values for modeling phenology, canopy development, and growth in the energy sorghum crop model used in this study.** The grain sorghum crop model parameters for a late maturity cultivar described in APSIM sorghum module (`apsim.info`) were modified to model energy sorghum. Energy sorghum phenology is approximated such that energy sorghum will initiate floral induction in mid-September given College Station, TX, environments as observed (Olson *et al.*, 2012; Rooney and Aydin, 1999). Other parameters were modified such that the model output generally resembled measurements of TX08001 in 2008 (Olson *et al.*, 2012).

Parameter description	CONSTANTS-SPECIES-SPECIFIC		
	Default	Adjusted	XML variable name
SECTION 1 - CROP PHENOLOGY: DEVELOPMENT PARAMETERS			
Maximum leaf number	40	50	leaf_no_max
Leaf appearance rate 1 (°Cd)	41	50	leaf_app_rate1
SECTION 2 - PHOTOSYNTHESIS BIOMASS GROWTH AND PARTITION			
Radiation use efficiency, RUE (g MJ-1) from juvenile to floral initiation	1.25	2.3	Rue
Extinction coefficient for green leaf from row spacing 0.5 m and 1.0 m	0.40	0.70	y_extinct_coef
SECTION 7 - SENESCENCE AND DETACHMENT			
Light delay factor for leaf senescence (days)	10.0	25.0	sen_light_time_const
Radiation level for onset of leaf senescence (MJ m-2)	2.0	0.5	sen_radn_crit
Water supply:demand ratio for onset of leaf senescence (MJ m-2)	0.25	0.03	sen_threshold
	GENOTYPE		
	late maturity	texas_energy	
	PHENOLOGY		
Photoperiod criteria 2	12.3	12.4	photoperiod_crit1
Photoperiod criteria 2	14.6	13.4	photoperiod_crit2
Photoperiod slope	38.6	1545	photoperiod_slope
	LEAF AREA - TPLA approach		
Curvature coefficient, α, for leaf area (1/°Cd)	0.018	0.003	tpla_prod_coef
Power coefficient, γ, for number of leaves to total plant leaf area	2.95	2.68	main_stem_coef
Inflection coefficient, β, of total plant leaf area curve	0.66	0.725	tpla_inflection_ratio
	CANOPY HEIGHT		
Minimum weight at maximum height to calculate density of stem accumulation (g/stem)	80	230	x_stem_wt
Maximum height (m)	2.0	4.0	y_height

calculate annual average ambient temperature (TAV) and annual amplitude in mean monthly temperature (AMP). The soil depth parameters were adjusted to be reflective of College Station, TX, where energy sorghum root systems rarely extend beyond 1000 mm below the soil surface. This constraint was implemented by making the 6th

soil layer (900-1200 mm depth) a water table that restricted root growth and water uptake past the 6th soil layer.

For management, sowing each year was modeled to occur between April 14th and May 1st. The sowing density and plant spacing in rows due to thinning reflect practices applied in College Station, TX, in 2008 and 2009 with 13.2 plants m^{-2} with 76 cm row spacing. Fertilization was simulated to apply 100 kg of nitrogen (N) per hectare on soil N demand, comparable with the production practices described in Olson *et al.* (2012). The unlimited irrigation regime was implemented using furrow irrigation between rows approximately every 2 weeks if no rainfall occurred and less often when rainfall provided water for plant growth (Olson *et al.*, 2012). For simulations without VPD-limited transpiration, irrigation of 150 mm of water was applied on soil water deficit demand of 50 mm as part of three irrigation scenarios: no irrigation (other than starting with a fully saturated soil profile prior to sowing, rainfall was the only water source), limited irrigation (irrigation stops on July 7th), and unlimited irrigation (irrigation on demand during the growing season).

4.3.2 Adapting APSIM sorghum module for the VPD-limited transpiration trait

In APSIM's sorghum model, biomass accumulation on a daily basis is calculated from the minimum of two potential biomass changes determined primarily from two inputs: radiation energy supply and soil water supply. For a given day, plant water demand is calculated from radiation energy supply, and the amount of water utilized by the plant during gas exchange required for biomass accumulation is calculated as the minimum of soil water supply and plant water demand. In other words, if water is not limiting, the amount of radiation received determines daily biomass accumulation; if water is limiting, the amount of water available determines daily biomass accumulation. To extend this model, VPD-limited transpiration was intro-

duced into the calculation of potential change in biomass per day by impacting plant water demand as a function of VPD. Daily weather input and biomass accumulation potential were interpolated at hourly timesteps, and for each daytime hour, the daily potential for biomass restricted by water was calculated as a function of hourly VPD. This effectively introduced the ability for a plant to slow its transpiration rate under conditions of high VPD (e.g., mid-afternoon), in which case the plant would demand less water under conditions of high VPD. Consequently, this slowed the rate of biomass accumulation under conditions of high VPD; less water was used in high VPD conditions where more water is necessary per unit of biomass accumulated. The calculations are described below, and their implementation in the APSIM sorghum C++ module is made available (see Code Availability).

VPD-limited transpiration was modeled using the parameters $\{vpd_{BP}, m_1, m_2\}$ which are typically obtained from experiments quantifying VPD-limited transpiration (Figure 4.2, Table 4.2). A VPD-limited transpiration trait can be characterized by a VPD breakpoint (vpd_{BP}), the slope of the linear relationship between transpiration rate and VPD for VPDs that are less than the VPD breakpoint (m_1), and the slope of the linear relationship between transpiration rate and VPD for VPDs that are greater than the VPD breakpoint (m_2); these three parameters can be defined in the .xml sorghum file provided by the user. Given these parameters, a daily biomass accumulation $\sum_{t=\text{sunrise}}^{\text{sunset}} B(t)$ is calculated on an hourly (t) basis over the course of a day length (sunrise to sunset). In order to evaluate VPD-limited transpiration on an hourly basis, VPD and biomass data calculated from radiation energy supply was interpolated hourly with respect to climate data by implementing the piecewise function described by Eccel (2010) from sunrise to sunset of sinusoid I and II equations. With this calculation, for each hour, t, there is an hourly transpiration rate, $T_r(t)$, based on potential biomass accumulation from radiation energy, $B_r(t)$, leaf

area index, lai, and an hourly vapor pressure deficit, $vpd(t)$, such that the following can be evaluated (Table 4.2 describes variables):

Transpiration rate dependent on radiation energy available,

$$[h!]T_r(t) = B_r(t) \div \frac{TE_c}{vpd(t)} \div lai \tag{4.1}$$

Transpiration rate dependent on VPD,

$$[h!]T_v(t) = \big((m_T(vpd(t)) \times vpd(t)) + (T(vpd_{BP}) - (m_T(vpd(t)) \times vpd_{BP}))\big), \tag{4.2}$$

where

$$[h!]m_T(vpd(t)) = \begin{cases} m_1, & vpd(t) < vpd_{BP} \\ m_2, & vpd(t) \geq vpd_{BP} \end{cases} \tag{4.3}$$

, and $T_v(t) \geq 0 \forall t$, and $T(vpd_{BP}) = m_1 \times vpd_{BP} - (-(1/4) \times m_1)$ so that at VPD less than or equal to $(1/4)$ there is no transpiration (Gholipoor *et al.*, 2010).

Transpiration rate dependent on VPD, radiation energy available, and extractable soil water,

$$[h!]T(t) = \min(T_v(t), T_r(t), (W_{soil} - W_{demand})). \tag{4.4}$$

Then, for every day time hour, daily soil water used is calculated as the sum of the hourly products of transpiration rate by leaf area, $T(t) \times lai$,

$$[h!]W_{demand}(t) = lai \times \sum_{i=t_{sunriset}}^{t} T(i) \tag{4.5}$$

And, daily biomass is calculated as the sum of the hourly product of transpiration

rate, $T(t)$, leaf area, lai, and transpiration efficiency, $\frac{TE_c}{vpd(t)}$.

$$[h!]B(t) = lai \times TE_c \times \sum_{i=t_{sunriset}}^{t} \frac{T(i)}{vpd(i)} \tag{4.6}$$

Table 4.2: **Descriptions of variables in the VPD-limited transpiration model modified within the APSIM sorghum module.** The time-scale column describes the time-scale (e.g. development stage, day, or hour) on which the variable changes or is evaluated. The origin column describes from where the variable is initialized or calculated (e.g. from a sub-class of the APSIM sorghum module, from the implementation of the VPD-limited transpiration model, or parameterized from experimental data).

Variable	**Description**	**Time-scale**	**Origin**
$B_r(t)$	Potential biomass from radiation energy interpolated from daily potential biomass calculation from radiation energy	Hour	Biomass class (APSIM)
lai	Leaf area index	Day	Leaf class (APSIM)
W_{soil}	Soil water available from water and root class	Day	Water class (APSIM)
TE_c	Genotype-specific, stage-dependent transpiration efficiency coefficient	developmental stage	Parameterized (Hammer *et al.*, 1997)
m_1	The slope of transpiration rate when $vpd(t)$ is less than vpd_{BP}.	Fixed	Parameterized
m_2	The slope of transpiration rate when $vpd(t)$ is greater than or equal to vpd_{BP}.	Fixed	Parameterized
vpd_{BP}	Vapor pressure deficit (VPD) at which transpiration rate transitions into another linear function.	Fixed	Parameterized
$vpd(t)$	Vapor pressure deficit (VPD) at hour t is interpolated from minimum and maximum daily temperature.	Hour	VPD-limited transpiration model
$T_r(t)$	Transpiration rate dependent of radiation energy available	Hour	VPD-limited transpiration model
$T_v(t)$	Transpiration rate dependent on vapor pressure deficit (VPD)	Hour	VPD-limited transpiration model
$T(t)$	Transpiration rate dependent on radiation energy available, VPD, and soil water available.	Hour	VPD-limited transpiration model
$W_{demand}(t)$	Cumulative soil water demand from plant from sunrise, tsunrise, to evaluated t. $W(t_{sunset})$ is the daily soil water demand that is passed back to the water class in APSIMs daily process.	Hour	VPD-limited transpiration model
$B(t)$	Cumulative biomass accumulated by plant from sunrise, tsunrise, to evaluated t. $B(t_{sunset})$ is the daily biomass that is passed back to the biomass class in APSIMs daily process.	Hour	VPD-limited transpiration model

4.3.3 Data and code accessibility

The files used to simulate energy sorghum crop growth and the code to modify growth based on VPD-limited transpiration can be found on GitHub (`www.github.com/MulletLab/sorghum-energy-crop-model/`).

4.4 Results

4.4.1 Modeling bioenergy sorghum in APSIM

To extend the applicability of APSIM to bioenergy sorghum, parameters of a sorghum genotype distributed with APSIM were modified to better reflect bioenergy sorghum characteristics, namely an extended period of vegetative growth, leaf area index (LAI), leaf number, radiation use efficiency, height, and stem density (Table Table 4.1). The sorghum model parameters in APSIM were adjusted based on energy sorghum traits measured as part of a previous field study (Olson *et al.*, 2012). This study characterized the growth and development of an energy sorghum hybrid, TX08001, near College Station, TX, in the 2008 and 2009 growing seasons that began in middle of April and ended in early November (Olson *et al.*, 2012). Three irrigation regimes were applied: a rain fed environment with no irrigation, a limited irrigation scheme where plants were watered as needed until July 7th (applied in 2008 and 2009) and an unlimited irrigation scheme where plants were watered throughout the season (applied in 2009). Above ground biomass was measured in both years, and in 2008, data on stem height, leaf number, and leaf area index was collected. After parameterization based on the limited irrigation 2008 trait data, APSIM simulations were compared with the remaining years and irrigation schemes.

Growth simulations of the energy sorghum for the unlimited, limited, and no irrigation conditions were qualitatively consistent with field data (Figure 4.1). In 2009, unlimited irrigation plots produced greater than 5 kg m^{-2} (50 Mg $hectare^{-1}$) of

dry shoot biomass (Figure 4.1A). In 2008 and 2009, limited irrigation plots produced approximately 4 kg m^{-2} (40 Mg $hectare^{-1}$) of dry shoot biomass (Figure 4.1D). The sorghum energy crop model qualitatively reproduced the growth trajectories observed in the unlimited and limited irrigation plots of 2008 and 2009. In addition to overall shoot biomass accumulation, the model also largely tracked LAI, height, and leaf number during the 2008 limited irrigation season (Figure 4.1F-H). Finally, in the targeted management regimes of energy sorghum, under strictly rain fed conditions, the bioenergy sorghum crop model predicted biomass accumulation that was within the range observed in 2008 mechanical harvests (Figure 4.1I). These results indicate that APSIM is capable of predicting growth trajectories observed in field experiments within the margin of uncertainty for bioenergy sorghum across multiple years and water supply conditions.

The parameterized APSIM energy sorghum model was next used to simulate how an energy sorghum crop would perform in the College Station, Texas, environment under different irrigation regimes and rainfall patterns over the annual growing seasons from 2000 to 2014 (Figure 4.1K). Using rainfall data from 2000 to 2014 in College Station, three water input regimes were imposed: unlimited, limited, and none (i.e. strictly rain fed). As expected, the crops ability to accumulate biomass in the three treatments diverges between 50-100 DAS when water becomes limiting, that is, when water in the initially saturated soil profile is depleted (Figure 4.1 B, E, J, K).

Fluctuations of environmental rainfall and VPD effect plant water use. Both rainfall and VPD environmental factors fluctuate on a seasonal basis, and VPD fluctuates significantly on a daily basis (Figure4.2). Sorghum transpiration efficiency (TE) is defined as a coefficient (9 Pa) over VPD (kPA), and so as VPD increases, TE decreases (Hammer *et al.*, 1997) (Figure 4.2C). The effect of high VPD on water use

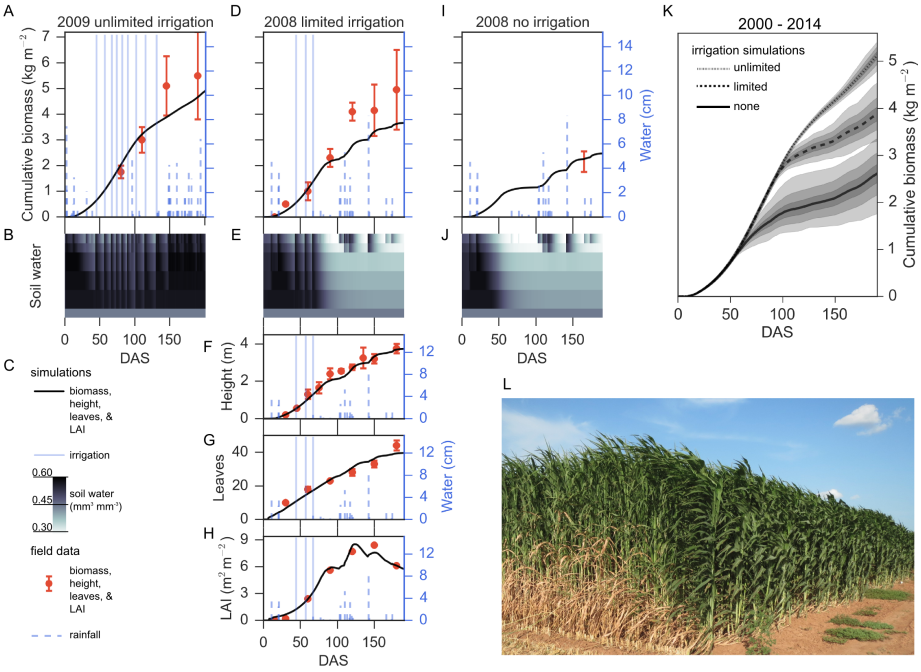

Figure 4.1: **Growth simulations and field data of energy sorghum in unlimited, limited, and no irrigation regimes.** Predicted biomass accumulation given 2009 and 2008 environmental conditions under unlimited (A), limited (D), and no (I) irrigation regimes fall within the margin of measurement error. Predicted leaf number (F), plant height (G), and leaf area index (LAI) (H) agree with experimental observations. Simulated water profiles from the three irrigation conditions are illustrated (B, E, J). Plotting descriptors are provided in panel C, and experimental data are plotted showing means and error bars representing one standard deviation where available, and the minimum and maximum range of observations in panel I Olson *et al.* (2012). (K) Average simulated biomass accumulation for the 2000 to 2014 cropping seasons for different irrigation regimes. The means of unlimited, limited, and no irrigation are plotted with continuous, dashed, and dotted lines, respectively. Each irrigation regime has 68, 98 and 100 percent confidence intervals depicted with decreasing shades of gray. The mean and confidence intervals were estimated with 50,000 bootstraps. (L) An image of energy sorghum hybrid, TX08001, in 2016 College Station, TX, cropping season 108 days after sowing.

is further compounded by the corresponding increase in water uptake from the soil profile by the root system as transpiration rate increases in response to increasing

VPD. That is, as VPD increases, the transpiration rate of sorghum increases, water uptake increases, but the efficiency of conversion of atmospheric carbon to plant sugars per unit of water transpired decreases (Figure 4.2C, D). In cases where water is not a limiting factor (and transpiration efficiency is not influencing biomass yield), energy sorghum is predicted to yield 5.4 ± 0.3 (mean ± SD) kg $^{-2}$ of biomass while using 1132 ± 82 (mean ± SD) mm $^{-2}$ of water during a 200 day growing season in simulations of unlimited irrigation of 2000 2014 College Station cropping environments. In contrast, in the absence of irrigation, the energy crop produced on average 2.7 ± 0.8 (mean ± SD) kg $^{-2}$ of biomass and uses 485 ± 120 (mean ± SD) mm $^{-2}$ of water (Figure 4.1K). These results indicate that water is a limiting factor of sorghum biomass accumulation.

4.4.2 Modeling VPD-limited transpiration trait

To examine the potential utility of the VPD-limited transpiration trait to improve water use in bioenergy sorghum, the APSIM sorghum growth model was modified. Since previous experimental work demonstrated that the VPD-limited transpiration trait is dependent on VPD and is sensitive to 1 kPA changes in VPD (Choudhary *et al.*, 2013; Choudhary and Sinclair, 2014; Gholipoor *et al.*, 2010; Riar *et al.*, 2015), and since the range of VPDs experienced by a crop over the course of a day is greater than 1 kPA due to daily variation in temperature, the APSIM model was modified so that the calculated daily water uptake could be altered by a VPD-limited transpiration trait calculated in hourly timesteps, an approach similar to that implemented by Sinclair *et al.* (2005); Messina *et al.* (2015) and Kholova *et al.* (2014).

VPD-limited transpiration trait was simplified into two different linear response slopes, m_1 and m_2, where the transition between m_1 and m_2 is determined by a

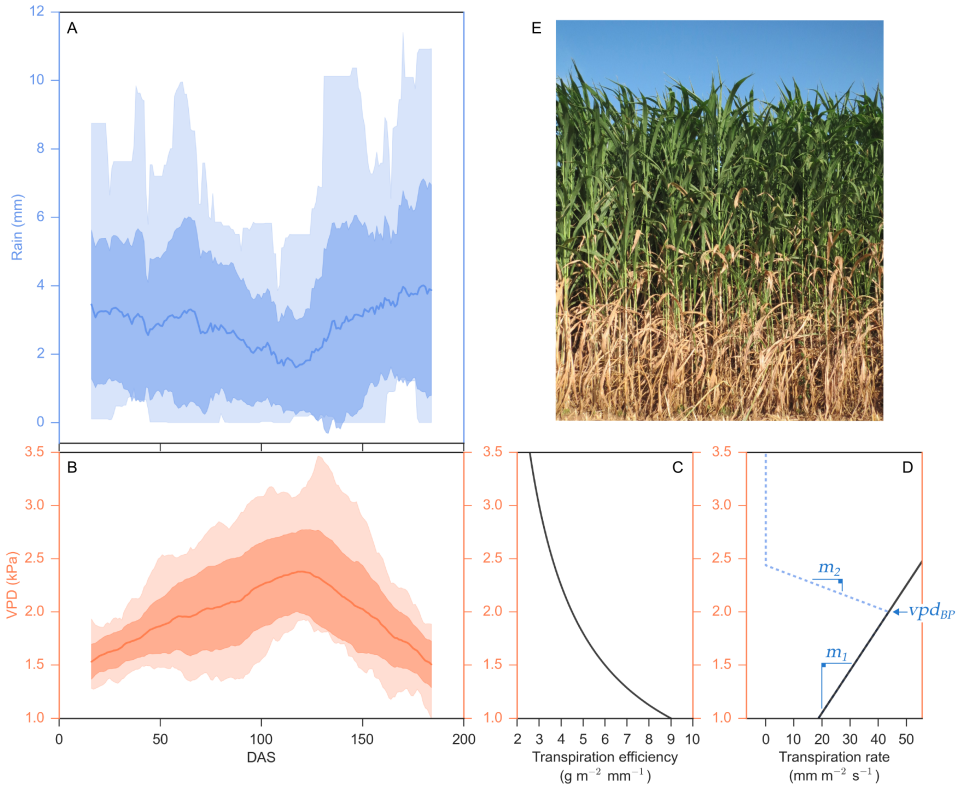

Figure 4.2: **Seasonal fluctuations of water and VPD and their relationship to sorghum transpiration rate and efficiency.** (A and B) Distribution of daily rainfall (mm) and daily vapor pressure deficit, VPD, (kPa) over the 2000-2014 cropping seasons in College Station, TX, calculated as a mean of a 30-day sliding window. The mean is plotted as a solid line, the lighter transparent fill is the entire data range (minimums and maximums) and the darker transparent fill is one standard deviation from the mean. (C) Transpiration efficiency, biomass produced per unit of water transpired (g m^{-2} mm^{-2}), and (D) a hypothetical transpiration rate (solid black line), the amount of water uptake per unit time (mm m^{-2} s^{-2}), and their responses to VPD (kPa^{-1}) are plotted along the x-axes sharing the y-axis of panel B. VPD-limited transpiration trait parameters $\{vpd_{BP}, m_1, m_2\}$ is denoted in blue. (E) Image of energy sorghum, TX08001, in College Station TX at 108 days after sowing in 2016.

VPD threshold, vpd_{BP} (Figure 2D). Such that, given genetic parameter set $\{m_1, m_2, vpd_{BP}\}$ and vpd(t), the VPD at hour t, transpiration rate, T(t), is described as the

following:

$$[h!]T(t) = \big((m_T(vpd(t)) \times vpd(t)) + (T(vpd_{BP}) - (m_T(vpd(t)) \times vpd_{BP}))\big), \quad (4.7)$$

where

$$[h!]m_T(vpd(t)) = \begin{cases} m_1, & vpd(t) < vpd_{BP} \\ m_2, & vpd(t) \geq vpd_{BP} \end{cases} \quad (4.8)$$

Specifically, VPD-limited transpiration using parameters m_1 , m_2 ,vpd_{BP} derived from studies of sorghum responses to VPD conducted by Choudhary *et al.* (2013); Choudhary and Sinclair (2014); Gholipoor *et al.* (2010) and, Riar *et al.* (2015) (Table 4.3 contains the parameters used). In this manner, each genotype is assigned VPD-limited transpiration parameters, $\{m_1$, m_2 ,$vpd_{BP}\}$ (denoted blue in Figure 4.2D), where m_1=m_2 if the plant has no unique vpd_{BP}, response of transpiration rate to VPD is linear (black line in Figure 4.2D).

Table 4.3: **Descriptions of variables in the VPD-limited transpiration model modified within the APSIM sorghum module.** The time-scale column describes the time-scale (e.g. development stage, day, or hour) on which the variable changes or is evaluated. The origin column describes from where the variable is initialized or calculated (e.g. from a sub-class of the APSIM sorghum module, from the implementation of the VPD-limited transpiration model, or parameterized from experimental data).

Genotype	m_1	m_2	vpd_{BP}	**Literature**
BQL41	20.26	7.27	1.17	Choudhary *et al.* (2013)
BTx623	56.3	10.4	2.05	Gholipoor *et al.* (2010)
SC803	39.8	-6.1	2.29	Gholipoor *et al.* (2010)
SC35	8.96	-12.49	2.91	Choudhary *et al.* (2013)
Tx436	6.62	-	-	Choudhary *et al.* (2013)

BQL41 is characterized by a particularly low vpd_{BP} of 1.17 kPA and its transpiration response to VPD is postulated to improve performance in dry environments (Choudhary *et al.*, 2013). To examine the physiological effect of BQL41s low vpd_{BP} on biomass yield and water use efficiency in water-limited environments, biomass accumulation of a VPD-limited transpiration trait that reflects the m_1 and vpd_{BP} of BQL41 were combined with variation in m_2 values. Five m_2 values were evaluated: (a) m_2=7.27, the transpiration rate that was experimentally obtained from the BQL41 genotype, (b) m_2=0.0, a maximum transpiration rate which reflects how VPD-limited transpiration has been modeled previously (Sinclair *et al.*, 2005; Messina *et al.*, 2015; Kholova *et al.*, 2014), (c) m_2=20.26, a response that reflects a lack of a vpd_{BP} (i.e., m_1=m_2=20.26), (d) m_2=-6.1, and (e) m_2=-12.49 transpiration responses to VPD that were experimentally obtained from SC803 and SC35 genotypes by Gholipoor *et al.* (2010) and Choudhary *et al.* (2013), respectively (Table 4.3).

The VPD-limited transpiration model was used to analyze the response of genotypes evaluated over the course of the growing season in the rain fed environments of College Station, TX (Figure 4.3). In these simulations, a less extreme m_2 of SC803 (m_2=-6.1) marginally outperformed the genotype lacking a vpd_{BP} over a 14 year average by 1.6 % (Supplemental Figure C.3, Supplemental Table C.1). The total end biomass accumulated by each of the genotypes investigated ranked from largest to smallest: 2.35 kg m^{-2} (SC803; m_2=-6.1), 2.31 kg m^{-2} (no vpd_{BP}; m_2=20.26), 2.29 kg m^{-2} (maximum limited transpiration; m_2=0), 2.28 kg m^{-2} (BQL41; m_2=7.27), and 2.25 kg m^{-2} (SC35; m_2=-12.49) (Supplemental Table C.1). The genotypes with a vpd_{BP} have slower growth and biomass accumulation at the beginning of the season, however as soil water begins to be depleted, the LAIs diverge such that genotypes with the lower m_2 have more live leaves than the genotypes with the higher m_2. This

biomass growth enabled by larger LAI is reflected in soil water extraction responses, where upon a period of rainfall later in the season, the genotypes with a vpd_{BP} are able to utilize more water and yield more biomass before the end of the season, like that of m_2=-6.1 (Supplemental Figure C.3).

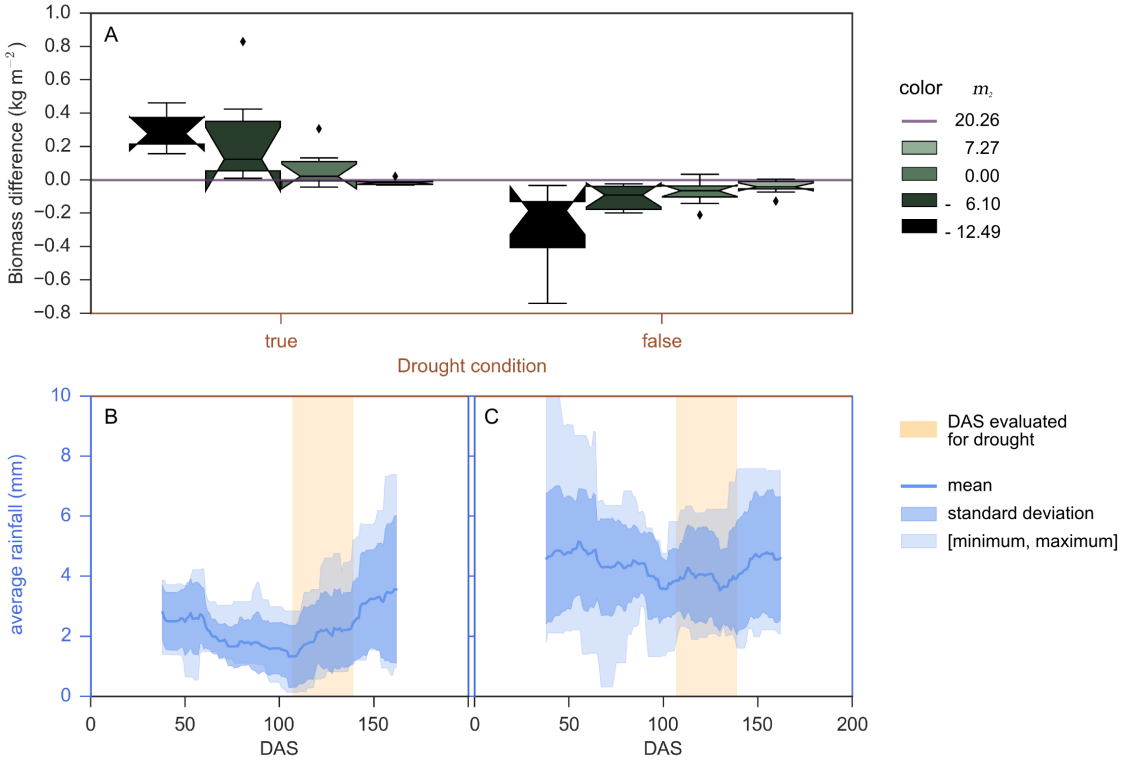

Figure 4.3: **Performance of VPD-limited transpiration determined by drought condition.** (A) Boxplots show the biomass difference between genotypes with a VPD-limited transpiration trait and the genotype that lacks a vpd_{BP} (m_2=20.26). The notches represent the 95% confidence interval of the median. These statistics are plotted against end biomass accumulated (DAS 200) by the genotype that lacks a vpd_{BP}, where the vertical sets of plots are separated by years categorized on the drought condition. Rainfall per day is calculated from sliding windows of 75 days. For sliding windows that center from DAS 107 139, if the average rainfall per day ever falls below 1.2 mm, then the year is considered a drought year. (B, C) The mean, standard deviation, and entire range of rainfall in sliding windows for the years considered drought and non-drought are plotted, respectively. The center of sliding windows evaluated for the drought condition is highlighted orange.

Further examination of biomass differences relative to timing of water availability suggests that the VPD-limited transpiration trait is beneficial when a long period of drought is experienced during the cropping season. A drought was defined as an event where the average rainfall per day fell below 1.2 mm given a 75-day sliding window average between DAS 107 139. Applying this drought criterion to rain-fed cropping environments of College Station, TX, for 2000 2014, and then evaluating the previous combination of m_1, m_2, and vpd_{BP} suggested that the VPD-limited transpiration trait is beneficial in water-limited environments, since VPD-limited transpiration plants with negative m_2 parameters always yielded greater biomass (Figure 4.3 and Supplemental Table C.1). Consistent with other crop models that predict limited transpiration plants perform better in conditions where non-limited transpiration plants will have low yield (Messina *et al.*, 2015), a sorghum genotype with maximum limited-transpiration (m_2=0) on average yielded more biomass than a sorghum genotype lacking a vpd_{BP} (m_2=m_1=20.26) in low yield environments (¡ 2.0 kg m^{-2} of the genotype lacking a vpd_{BP} in rain fed conditions of College Station, TX 2000 2014) (Supplemental Figure C.4). Moreover, differences in predicted canopy maintenance are also an effective predictor of years where VPD-limited transpiration plants outperform their non-limited transpiration counterparts (Supplemental Figure C.5), a property that has been observed in other limited-transpiration crop models (Hammer *et al.*, 2016; Kholova *et al.*, 2014). Regardless of the criteria used to predict conditions where VPD-limited transpiration will be beneficial (i.e. drought, yield, or LAI), more restrictive transpiration responses associated with more negative m_2 values perform progressively better in water-limited, low-yielding environments that lead to loss of canopy, but progressively worse in higher-yielding conditions where crop canopies are maintained. This resulting trade-off is further emphasized in a simulated terminal drought versus well-watered condition (Supplemental Information

and Supplemental Figure C.6), where water availability dictates whether the trait is beneficial or a liability.

To investigate the impact of variation in m_1 slopes, plants lacking a vpd_{BP} were simulated to reflect the range of transpiration responses to varying VPD that have been experimentally identified ($m_1 \in \{6.62, 56.3\}$) and a hypothetical extreme (m_1=200.0). Specifically, TX436 (m_1= 6.62; Choudhary *et al.* (2013)), BQL41 (m_1=20.26; Choudhary *et al.* (2013)), and BTX623 (m_1= 56.3; Gholipoor *et al.* (2010)) and the hypothetical extreme (m_1=200.0) were simulated using the VPD-limited transpiration modification with no vpd_{BP} (m_1=m_2) (Figure 4.4). On average across 14 years in rain fed College Station, TX, cropping environments, increased values of m_1 resulted in higher biomass accumulation, with m_1 of 56.3, yielding approximately 2.5 kg m^{-2}, similar to the daily model without VPD-limited transpiration model. Higher m_1 values (m_1¿56.3) did not increase end-point biomass accumulation. While biomass accumulation was positively related to larger m_1 values, it also resulted in slower canopy development (LAI) and greater leaf senescence (Figure 4.4D). Despite reductions in leaf area, genotypes with larger m_1 values still produced more biomass than genotypes with smaller m_1 values on average over the 14 years analyzed.

To examine the influence of all three parameters describing VPD-limited transpiration, vpd_{BP} and m_2 were varied within the genetic bounds observed in the literature, and their impact on biomass yields given different m_1 parameters were evaluated ($m_1$6.62, 20.26, 56.3). A range of vpd_{BP} values ($vpd_{BP} \in \{1.17, 2.91\}$) representing the range bounded by BQL41 and SC35, respectively, and a range of m_2 values ($m_2 \in \{$-12.49, 56.3$\}$) representing the range bounded by SC35 and BTX623, respectively, were explored (Gholipoor *et al.*, 2010; Choudhary *et al.*, 2013). Combinations of these parameters were simulated in rain fed College Station, TX, en-

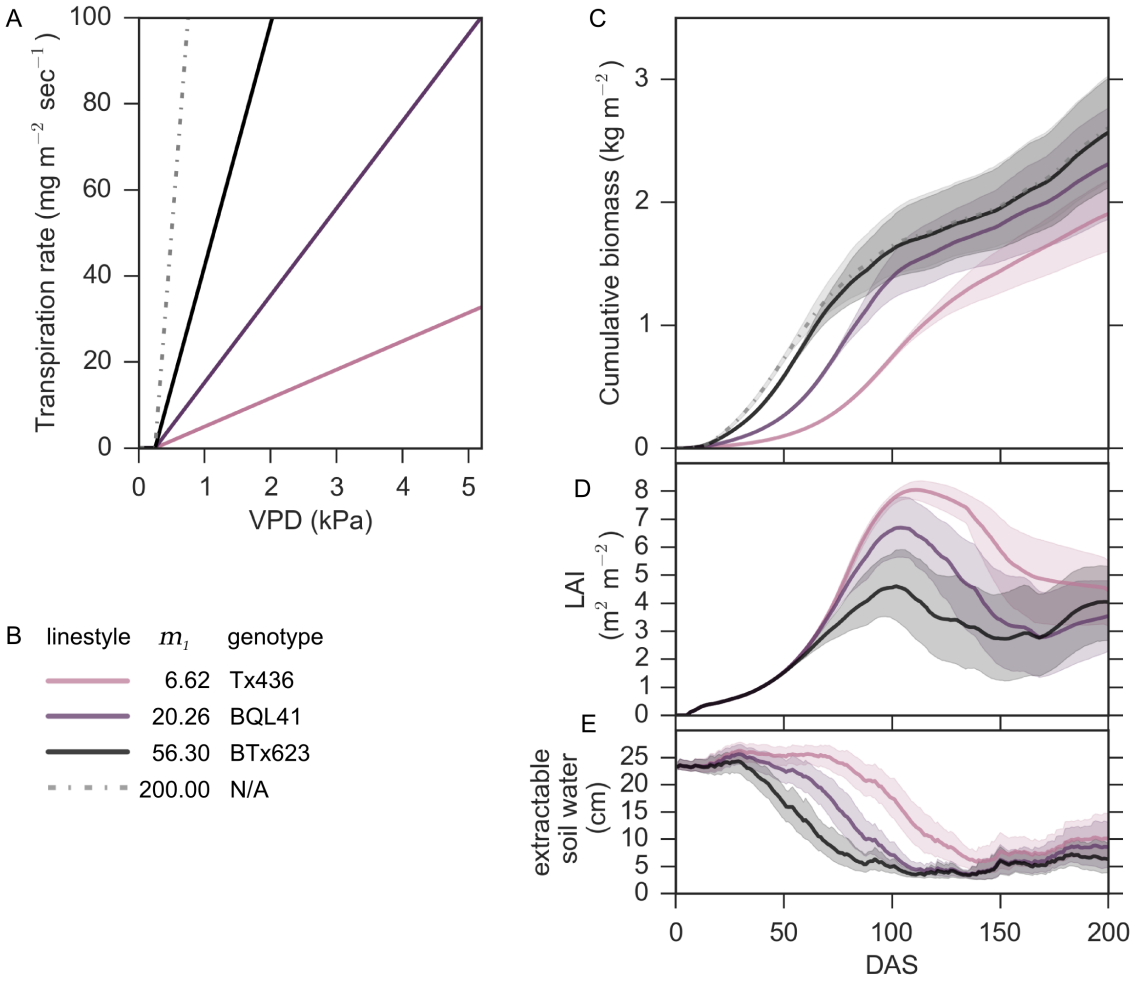

Figure 4.4: **Energy sorghum biomass accumulation from 2000-2014 cropping seasons simulated in College Station, TX, with water supply coming only from rainfall.** (A-D) Growth simulation of plants with no vpd_{BP} (m_1= m_2) are plotted with continuous colored lines that correspond to their $m_1$6.62, 20.26, 56.3, 200. (C) Cumulative biomass, (D) leaf area index (LAI), and (E) extractable soil water are plotted as the means and 98% confidence intervals estimated with 50,000 bootstraps.

vironments of 2000 2014 cropping seasons. The total biomass accumulated at 200 days after sowing (DAS) is used as a measurement of crop productivity. These productivity landscapes illustrate the influence of VPD-limited transpiration trait, and

indicate that, on average, in the target environment of College Station, TX, m_1 is the greatest driver of average productivity of the sorghum crop (Figure 4.5). All simulated genotypes with an m_1=56.3 produced more biomass than those with an m_1=20.26 or 6.62, regardless of m_2 and vpd_{BP}.

Contour plots of the three m_1s show that the extent of phenotypic variation introduced by the other two parameters m_2 and vpd_{BP} is influenced by the magnitude of m_1, where the variation in cumulative biomass increases as m_1 decreases. Furthermore, from analysis of BQL41, the relationship between transpiration rates m_1 and m_2, where the beneficial changes were observed when where $m_2 \cong -m_1/3$, which is not within the genetic parameter space observed given an m_1=56.3. This suggested that there may exist a theoretical m_2 for m_1=56.3 that would be predicted to increase productivity beyond the 0.2% in the observed parameter space.

For an m_1=56.3, the exploration of a theoretical m_2 outside what has been previously observed for VPD-limited-transpiration predicted that an m_2= -290 with a vpd_{BP}= 2.05 would increase biomass by 6.5%, yielding an average of 2.73 kg m^{-2} in rain fed College Station, TX (2000-2014) (Figure 4.6). This extreme m_2= -290 would be reflective of a genotype that essentially stops transpiration at VPDs higher than its vpd_{BP}. Furthermore, observation of the fitness landscape suggests that it is not essential to find the exact combination of vpd_{BP} and m_2, but that there is a gradual ridge constructed by combinations of the two VPD-limited transpiration parameters that would improve biomass accumulation by $\geq$5% (Figure 6). This ridge shows the relationship between the vpd_{BP} and m_2 for increasing biomass by 5% in rain fed College Station, TX (2000 2014) in the VPD-limited transpiration diagram, where as vpd_{BP} increased, the range of m_2 parameters to combine for beneficial productivity increases. This relationship with vpd_{BP} is likely related to the effect of VPD on transpiration efficiency, TE, which is inversely proportional to VPD through the

TE_c coefficient: $TE = \frac{TE_c}{vpd}$.

4.5 Discussion

The bioenergy sorghum crop model leverages the utility of the well-developed APSIM framework and the APSIM-sorghum crop model and is made available for evaluation in target environments so that general production predictions can be evaluated in silico to inform experimental design or production deployment. Evaluation in College Station, TX, serves as an example of a relatively hot and dry environment where bioenergy sorghum is targeted to be deployed. Under rain fed conditions in College Station, Texas, where the median rainfall over the growing season used for the simulations was 52.3 cm, bioenergy sorghum is projected to yield approximately 25 Mg hectare^{-1} per year. This model prediction is consistent with the biomass yield of TX08001 ($\sim$15-26 Mg hectare^{-1}) grown without irrigation in large plots that were machine harvested in 2008 and 2009 (Olson *et al.*, 2012), suggesting that the bioenergy crop model can be used to evaluate in future target environments.

Adapting the bioenergy sorghum model to evaluate the VPD-limited transpiration trait indicated that certain VPD-limited transpiration characteristics can be beneficial for bioenergy sorghum in water-limited rain-fed environments, providing a potential means to increase biomass production on annual cropland that is marginal for many food crops. The available model extension can be used to evaluate whether the trait would be desirable in other the target production environments and to determine trade-offs to be considered, as indicated by the College Station, TX, case study (mean $\pm$ SD: 56 $\pm$ 18 cm of rainfall per cropping season for 2000 2014). The VPD-limited transpiration trait benefits plants most when increased transpiration efficiency can compensate for reduced growth rate. However, this improvement is not attained if water is not sufficiently limited, an emergent property that has

been associated with limited transpiration traits in grain sorghum (Hammer et al., 2016; Kholov et al., 2014). In general, plants with smaller m_1 or m_2 values transpire less water per day, resulting in lower water demand, and consequently, these plants showed delayed and reduced leaf senescence during the water-limited portion of the season. This is beneficial if leaf area would otherwise dramatically fall due to water stress. Examining the rainfall distribution and its effect on VPD-limited transpiration trait showed that in drought conditions m_2=-12.49 gained 20% more biomass than genotypes lacking a vpd_{BP} and in non-drought conditions generate 10% less biomass, illustrating this trade-off (Supplemental Table C.1). Determining target environments that benefit from VPD-limited transpiration will be critical in testing the utility of the trait as well as in screening for energy sorghum hybrids that exhibit VPD-limited transpiration (Tardieu, 2011).

We note that VPD-limited transpiration given the constraints of parameters from previously studied germplasm provides little in the way of average yield benefits in College Station, TX, environment. However significant benefits are obtainable outside of these parameter constraints (i.e. m_2¡-200). Evidence by work from others suggests that limiting transpiration can be environmentally triggered by changes in VPD (Lobet *et al.*, 2014; McAdam and Brodribb, 2015; Kholova *et al.*, 2014), by driving foliar accumulation of the phytohormone abscisic acid (ABA) (Assmann *et al.*, 2000; Bauer *et al.*, 2013; McAdam and Brodribb, 2015; McAdam *et al.*, 2015), and the increase of ABA modifies stomatal aperture in turn reducing plant transpiration (Boyer, 2010; Cutler *et al.*, 2010; Parent *et al.*, 2009; Tardieu and Davies, 1992). Stomatal closure can be further engineered via the ABA pathway, whereby application of agrochemicals on plants expressing genetically engineered ABA drastically reduces transpiration on demand (Park *et al.*, 2015). Genetic engineering approaches like these may provide a means to design plants with more negative m_2 and improved

yields.

In conclusion, this study (i) demonstrated that modeling of energy sorghum in the APSIM framework is able to track field growth in different environments, (ii) extended the daily APSIM-sorghum model to incorporate hourly VPD-limited transpiration, (iii) found that VPD-limited transpiration can improve crop productivity in water-limiting environments by maintenance of crop canopies to utilize sporadic rainfall, (iv) and make this model and extension available for further evaluation of bioenergy sorghum production in other targeted environments.

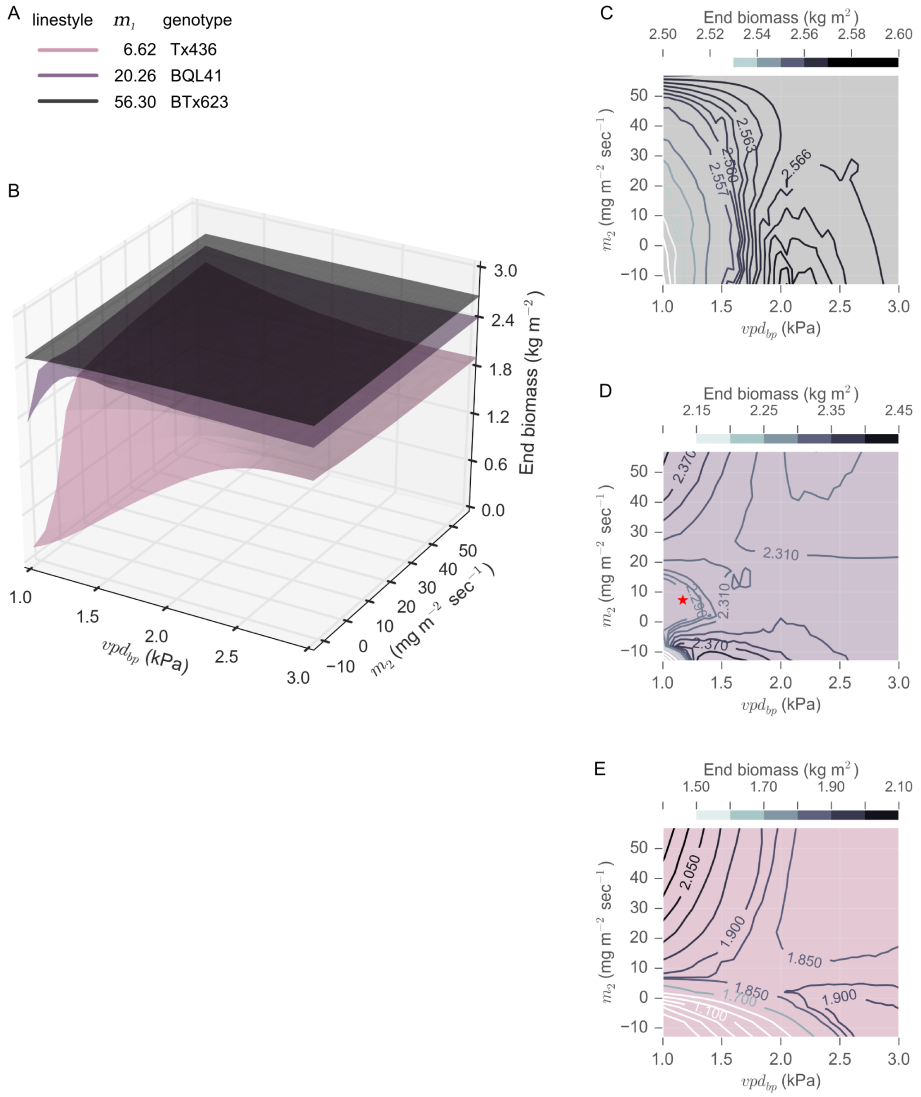

Figure 4.5: **Landscapes of energy sorghum productivity for varying m_1, m_2 and vpd_{BP} within the observed phenotypic parameters.** End biomass at 200 DAS predicted from rain fed College Station, TX (2000 2014) was used as a measure of productivity. (A) m_1 was varied to reflect the extremes (6.62, Tx426 and 56.3, BTx623) and a moderate transpiration rate (20.26, BQL41). (B) $m_2 \in$ {-12.49, 56.3} and $vpd_{BP} \in \{1.17, 2.91\}$ are evaluated for each m_1 and the predicted average end biomass was interpolated to construct the topological surfaces. (D-F) The topology of each m_1 surface is also plotted two-dimensionally, with contour lines to depict variation of end biomass at 200 DAS. The red star in panel E represents the VPD-limited transpiration parameters of BQL41.

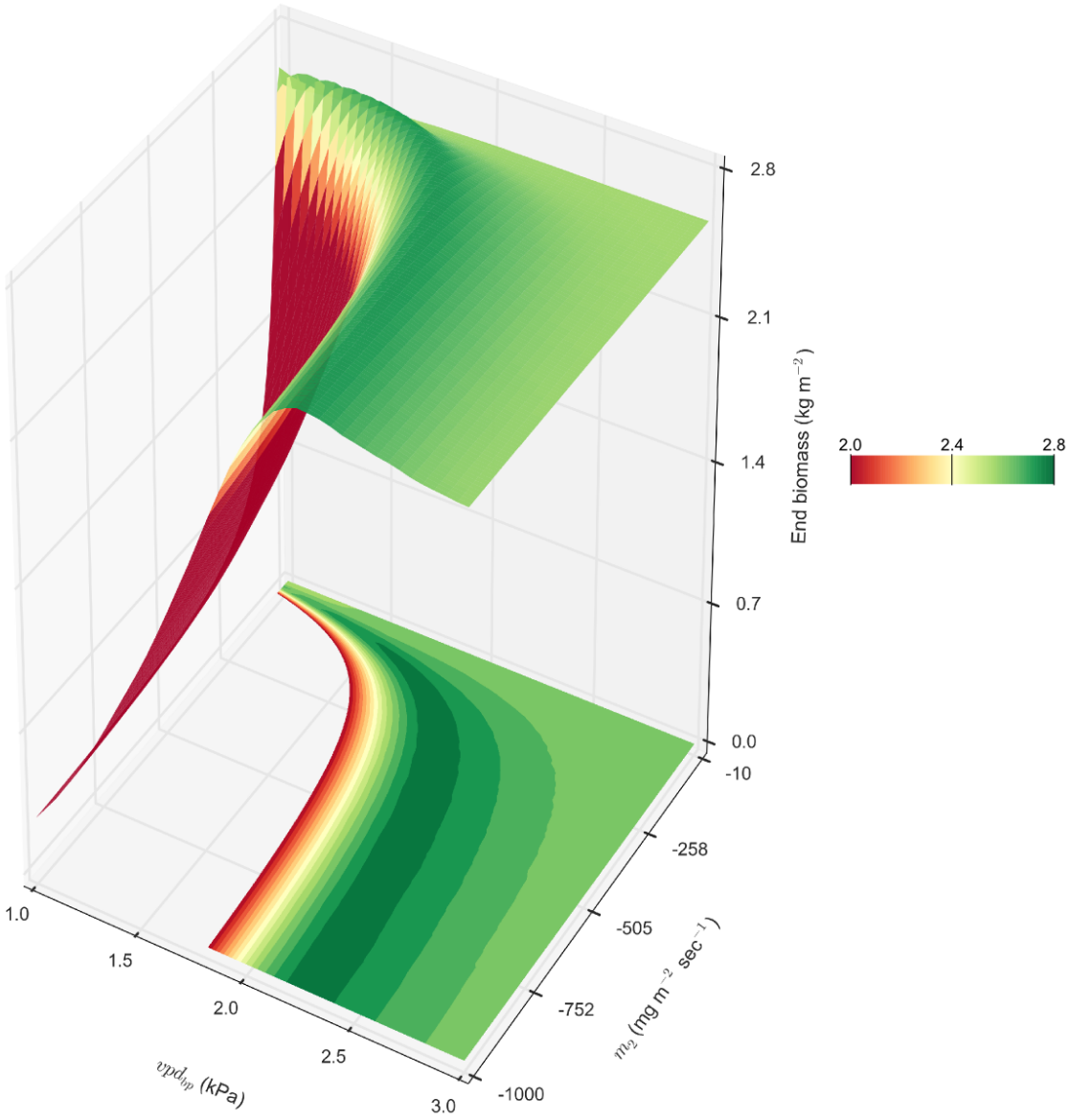

Figure 4.6: **Theoretical energy sorghum productivity landscape.** (A) A combination of m_1=56.3 representing BTx623, a range of vpd_{BP} values (vpd_{BP} ∈{1.17, 2.91}) representing BQL41 and SC35, and a range of m_2 values (m_2 ∈{-1000,-10}) beyond the observed genotypic parameters were evaluated for end biomass in rain fed College Station, TX, cropping environments 2000 2014, and their averages are interpolated to create the topology of end productivity.

5. CONCLUSION

Crop production will need to approximately double by 2050 to meet the global projected caloric and energy demands from imposed by the increasing size and economic affluence of the global population (Godfray *et al.*, 2010; Alexandratos and Bruinsma, 2012; Foley *et al.*, 2011; Tilman *et al.*, 2011). However, these demands cannot be sustainably met given the current trends in yield improvement. Agricultural intensification must shift its methods to mitigate damage to to natural ecosystems and environment, and be implemented in the context of projected conditions, where anthropogenic climate change is causing increased uncertainty for plant productivity (Foley *et al.*, 2011; Mueller *et al.*, 2012; Tilman *et al.*, 2011; Porter *et al.*, 2010; Lobell *et al.*, 2014).

Towards the goal of sustainably increasing crop productivity, this dissertation describes the use of modeling and genetics to study sorghum traits that yield performance improvements in current and projected production environments. Chapter 2 developed a model to improve calculations of genetic linkage in plant recombinant inbred line populations so that regions of the genome associated with phenotypic variation can be more accurately defined. Chapter 3 constructed functional-structural plant models to explore the effect of a crop canopy architecture trait, leaf angle, on plant productivity. Combined with field experimentation and genetic analyses this lead to the identification of genomic loci regulating leaf angle and quantified its predicted impact on biomass accumulation. Finally, chapter 4 parameterized a bioenergy sorghum crop model to predict production of lignocellulosic biomass and was extended to explore how VPD-limited transpiration influenced the crop growth and end productivity through its effect of plant water use. This work contributes

towards the design and identification of robust and superior sorghum ideotypes for target production environments.

Improving crop productivity is currently hindered by the rate at which we can alter biological systems with predictable consequences. Research into crop improvement is composed of many disciplines, from biochemistry and molecular genetics to economic and climate analyses, and interrelating knowledge across disciplines will improve the rate with which innovative solutions can be employed to provide more efficient food and economically competitive energy sources. The integration of mathematical and statistical models can be leveraged to identify emergent properties in crop production systems to complement and prioritize experiments that improve economic gains. As an example, genomic prediction models could be used to predict the allelic effects of genetic variants on quantitative traits that then are used to parameterize functional-structural plant models to simulate microenvironments such as light distribution in the crop canopy. The effects of the redistribution of light can then be used in crop-level models to predict overall crop productivity, and yield information can be used for regional economic modeling given current and projected climate environments. From these models, productivity landscapes can be searched to identify optimal allelic combinations that can then be designed and deployed in multi-site field trials for further testing. In this manner, the phenotypic search space can be more rapidly and intelligently searched to increase the rate of yield gains in agricultural crops. Moreover specific work into integrating or interrelating these ideas through modeling has shown promise from industry and academia (Technow *et al.*, 2015; Cooper *et al.*, 2016; McDowell *et al.*, 2016).

Integrating and interrelating tools and methods from multiple perspectives has accelerated research, and there will be continued gains from interdisciplinary work to generate high-performing cultivars in target current and projected environments.

REFERENCES

Alexandratos, N., and J. Bruinsma, 2012 World agriculture towards 2030/2050: the 2012 revision. ESA Working paper **12-03**.

Anderson, L. K., G. G. Doyle, B. Brigham, J. Carter, K. D. Hooker, *et al.*, 2003 High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. Genetics **165**: 849–865.

Assmann, S. M., J. A. Snyder, and Y. R. J. Lee, 2000 Aba-deficient (*aba1*) and aba-insensitive (*abi1-1*, *abi2-1*) mutants of arabidopsis have a wild-type stomatal response to humidity. Plant Cell and Environment **23**: 387–395.

Barrero Farfan, I. D., B. R. Bergsma, G. Johal, and M. R. Tuinstra, 2012 A stable allele in sorghum and a molecular marker to facilitate selection. Crop Science **52**: 2063–2069.

Bartek, M. S., S. Murray, P. Klein, J. Mullet, and W. Rooney, 2012 QTL for biomass yield and composition in energy sorghum (*Sorghum bicolor* L. Moench). Science for Biomass Feedstock Production and Utilization SunGrant Conference, New Orleans, LA, USA .

Bauer, H., P. Ache, S. Lautner, J. Fromm, W. Hartung, *et al.*, 2013 The stomatal response to reduced relative humidity requires guard cell-autonomous aba synthesis. Current Biology **23**: 53–57.

Beavis, W., 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the forty-ninth annual corn and sorghum industry research conference*. Chicago, IL, 250–266.

Beavis, W., 1998 QTL analyses: power, precision, and accuracy. In A. Patterson, M. Baker and C. Whitehead, editors, *Molecular dissection of complex traits*, chap-

ter 10. CRC Press, 145–162.

Bhattramakki, D., J. Dong, A. K. Chhabra, and G. E. Hart, 2000 An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) moench. Genome **43**: 988–1002.

Birch, C., P. Carberry, R. Muchow, R. McCown, and J. Hargreaves, 1990 Development and evaluation of a sorghum model based on ceres-maize in a semi-arid tropical environment. Field Crops Research **24**: 87–104.

Birch, C. J., J. Vos, J. Kiniry, H. J. Bos, and A. Elings, 1998 Phyllochron responds to acclimation to temperature and irradiance in maize. Field Crops Research **59**: 187–200.

Blakeslee, J. J., A. Bandyopadhyay, O. R. Lee, J. Mravec, B. Titapiwatanakun, *et al.*, 2007 Interactions among pin-formed and p-glycoprotein auxin transporters in arabidopsis. The Plant Cell **19**: 131–147.

Blum, A., 2004 Sorghum physiology. In H. T. Nguyen and A. Blum, editors, *Physiology and biotechnology integration for plant breeding*. CRC Press, 141–224.

Borrell, A. K., G. L. Hammer, and A. C. L. Douglas, 2000a Does maintaining green leaf area in sorghum improve yield under drought? i. leaf growth and senescence. Crop Science **40**: 1026–1037.

Borrell, A. K., G. L. Hammer, and R. G. Henzell, 2000b Does maintaining green leaf area in sorghum improve yield under drought? ii. dry matter production and yield. Crop Science **40**: 1037–1048.

Borrell, A. K., J. E. Mullet, B. George-Jaeggli, E. J. van Oosterom, G. L. Hammer, *et al.*, 2014 Drought adaptation of stay-green sorghum is associated with canopy development, leaf anatomy, root growth, and water uptake. Journal of Experimental Botany **65**: 6251–6263.

Boudon, F., C. Pradal, T. Cokelaer, P. Prusinkiewicz, and C. Godin, 2012 L-py: an l-system simulation framework for modeling plant architecture development based

on a dynamic language. Frontiers in Plant Science **3**: 76.

Boyer, J. S., 2010 Drought decision-making. Journal of Experimental Botany **61**: 3493–3497.

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber, 1998 Comprehensive human genetic maps: Individual and sex-specific variation in recombination. The American Journal of Human Genetics **63**: 861 – 869.

Broman, K. W., and J. L. Weber, 2000 Characterization of human crossover interference. The American Journal of Human Genetics **66**: 1911–1926.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**: 889–890.

Bulmer, M. G., *et al.*, 1980 *The mathematical theory of quantitative genetics*. Clarendon Press.

Burow, G., R. Klein, C. Franks, P. Klein, K. Schertz, *et al.*, 2011 Registration of the BTx623/IS3620C recombinant inbred mapping population of sorghum. Journal of Plant Registrations **5**: 141–145.

Burr, B., and F. A. Burr, 1991 Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations. Trends in Genetics **7**: 55–60.

Chapman, S., M. Cooper, G. Hammer, and D. Butler, 2000a Genotype by environment interactions affecting grain sorghum. ii. frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Crop and Pasture Science **51**: 209–222.

Chapman, S., G. Hammer, D. Butler, and M. Cooper, 2000b Genotype by environment interactions affecting grain sorghum. iii. temporal sequences and spatial patterns in the target population of environments. Crop and Pasture Science **51**: 223–234.

Chelle, M., and B. Andrieu, 1998 The nested radiosity model for the distribution of

light within plant canopies. Ecological Modelling **111**: 75–91.

Chelle, M., J. Hanan, and H. Autret, 2004 Lighting virtual crops: the caribu solution for open l-systems. In *4th International Workshop on Functional-Structural Plant Models, Montpellier*.

Choudhary, S., R. N. Mutava, A. Shekoofa, T. R. Sinclair, and P. Prasad, 2013 Is the stay-green trait in sorghum a result of transpiration sensitivity to either soil drying or vapor pressure deficit? Crop Science **53**: 2129–2134.

Choudhary, S., and T. R. Sinclair, 2014 Hydraulic conductance differences among sorghum genotypes to explain variation in restricted transpiration rates. Functional Plant Biology **41**: 270–275.

Clerget, B., M. Dingkuhn, E. Gozé, H. Rattunde, and B. Ney, 2008 Variability of phyllochron, plastochron and rate of increase in height in photoperiod-sensitive sorghum varieties. Annals of Botany **101**: 579–594.

Cloutier, S., R. Ragupathy, E. Miranda, N. Radovanovic, E. Reimer, *et al.*, 2012 Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). Theoretical and Applied Genetics **125**: 1783–1795.

Cooper, M., F. Technow, C. Messina, C. Gho, and L. R. Totir, 2016 Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. Crop Science **56**: 2141–2156.

Cutler, S. R., P. L. Rodriguez, R. R. Finkelstein, and S. R. Abrams, 2010 Abscisic acid: Emergence of a core signaling network. Annual Review of Plant Biology **61**: 651–679.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics **43**: 491–498.

Dib, C., S. Fauré, C. Fizames, D. Samson, N. Drouot, *et al.*, 1996 A comprehensive

genetic map of the human genome based on 5,264 microsatellites. Nature **380**: 152–154.

Doggett, H., 1988 *Sorghum*. Harlow, Essex, England New York, 2nd edition.

Dong, Z., C. Jiang, X. Chen, T. Zhang, L. Ding, *et al.*, 2013 Maize LAZY1 mediates shoot gravitropism and inflorescence development through regulating auxin transport, auxin signaling, and light response. Plant Physiology **163**: 1306–1322.

Drewry, D. T., P. Kumar, and S. P. Long, 2014 Simultaneous improvement in productivity, water use, and albedo through crop structural modification. Global Change Biology **20**: 1955–1967.

Drouet, J.-L., and R. Bonhomme, 1999 Do variations in local leaf irradiance explain changes to leaf nitrogen within row maize canopies? Annals of Botany **84**: 61–69.

Duvick, D., 2005a Genetic progress in yield of united states maize (*Zea mays* l.). Maydica **50**: 193.

Duvick, D. N., 2005b Genetic progress in yield of United States maize (*Zea mays* L.). Maydica **50**: 193–202.

Eccel, E., 2010 What we can ask to hourly temperature recording. part ii: hourly interpolation of temperatures for climatology and modelling. Italian Journal of Agrometerology - Rivista Italiana Di Agormeteorologia **15**: 45–50.

Foley, J. A., N. Ramankutty, K. A. Brauman, E. S. Cassidy, J. S. Gerber, *et al.*, 2011 Solutions for a cultivated planet. Nature **478**: 337–342.

George-Jaeggli, B., D. Jordan, E. Van Oosterom, I. Broad, and G. Hammer, 2013 Sorghum dwarfing genes can affect radiation capture and radiation use efficiency. Field Crops Research **149**: 283–290.

George-Jaeggli, B., D. Jordan, E. Van Oosterom, and G. Hammer, 2011 Decrease in sorghum grain yield due to the *dw3* dwarfing gene is caused by reduction in shoot biomass. Field Crops Research **124**: 231–239.

Gholipoor, M., P. V. Prasad, R. N. Mutava, and T. R. Sinclair, 2010 Genetic variability of transpiration response to vapor pressure deficit among sorghum genotypes. Field Crops Research **119**: 85–90.

Gill, J. R., P. S. Burks, S. A. Staggenborg, G. N. Odvody, R. W. Heiniger, *et al.*, 2014 Yield results and stability analysis from the sorghum regional biomass feedstock trial. Bioenergy Research **7**: 1026–1034.

Godfray, H. C. J., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, *et al.*, 2010 Food security: the challenge of feeding 9 billion people. Science **327**: 812–818.

Haldane, J., and C. Waddington, 1931 Inbreeding and linkage. Genetics **16**: 357.

Hall, K. J., J. S. Parker, and T. H. N. Ellis, 1997a The relationship between genetic and cytogenetic maps of pea. I. Standard and translocation karyotypes. Genome **40**: 744–754.

Hall, K. J., J. S. Parker, T. H. N. Ellis, L. Turner, M. R. Knox, *et al.*, 1997b The relationship between genetic and cytogenetic maps of pea. II. Physical maps of linkage mapping populations. Genome **40**: 755–769.

Hammer, G., G. McLean, A. Doherty, E. van Oosterom, and S. Chapman, 2016 Sorghum crop modeling and its utility in agronomy and breeding. Sorghum: State of the Art and Future Perspectives .

Hammer, G., and R. C. Muchow, 1994 Assessing climatic risk to sorghum production in water-limited subtropical environments i. development and testing of a simulation model. Field Crops Research **36**: 221–234.

Hammer, G. L., Z. Dong, G. McLean, A. Doherty, C. Messina, *et al.*, 2009 Can changes in canopy and/or root system architecture explain historical maize yield trends in the us corn belt? Crop Science **49**: 299–312.

Hammer, G. L., G. D. Farquhar, and I. J. Broad, 1997 On the extent of genetic variation for transpiration efficiency in sorghum. Australian Journal of Agricultural

Research **48**: 649–655.

Hammer, G. L., E. van Oosterom, G. McLean, S. C. Chapman, I. Broad, *et al.*, 2010 Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. Journal of Experimental Botany : erq095.

Hart, G. E., K. F. Schertz, Y. Peng, and N. H. Syed, 2001 Genetic mapping of *Sorghum bicolor* (L.) Moench QTLs that control variation in tillering and other morphological characters. Theoretical and Applied Genetics **103**: 1232–1242.

Hilley, J., S. Truong, S. Olson, D. Morishige, and J. Mullet, 2016 Identification of *Dw1*, a regulator of sorghum stem internode length. PloS one **11**: e0151271.

Holzworth, D. P., N. I. Huth, E. J. Zurcher, N. I. Herrmann, G. McLean, *et al.*, 2014 Apsimevolution towards a new generation of agricultural systems simulation. Environmental Modelling and Software **62**: 327–350.

Hurng, W., L. Su, and C. Kao, 1986 Senescence of rice leaves. 16. regulation by light. Botanical Bulletin of Academia Sinica **27**: 163–174.

Jones, E., O. T. P. P. e. a., 2001 Scipy: Open source scientific tools for python.

Keating, B. A., P. S. Carberry, G. L. Hammer, M. E. Probert, M. J. Robertson, *et al.*, 2003 An overview of apsim, a model designed for farming systems simulation. European Journal of Agronomy **18**: 267–288.

Kholova, J., T. Murugesan, S. Kaliamoorthy, S. Malayee, R. Baddam, *et al.*, 2014 Modelling the effect of plant water use traits on yield and stay-green expression in sorghum. Functional Plant Biology **41**: 1019–1034.

Kimber, C., 2000 Origins of domesticated sorghum and its early diffusion to India and China. Sorghum: Origin, history, technology, and production : 3–98.

King, J., L. A. Roberts, M. J. Kearsey, H. M. Thomas, R. N. Jones, *et al.*, 2002 A demonstration of a 1:1 correspondence between chiasma frequency and recombination using a *Lolium perenne*/*Festuca pratensis* substitution. Genetics **161**:

307–314.

Knöller, A. S., J. J. Blakeslee, E. L. Richards, W. A. Peer, and A. S. Murphy, 2010 *Brachytic2/ZmABCB1* functions in IAA export from intercalary meristems. Journal of Experimental Botany **61**: 3689–3696.

Knox, M. R., and T. H. N. Ellis, 2002 Excess heterozygosity contributes to genetic map expansion in pea recombinant inbred populations. Genetics **162**: 861–873.

Kong, W., H. Jin, C. D. Franks, C. Kim, R. Bandopadhyay, *et al.*, 2013 Genetic analysis of recombinant inbred lines for *Sorghum bicolor* × *Sorghum propinquum*. G3: Genes— Genomes— Genetics **3**: 101–108.

Kumar, S. R., G. L. Hammer, I. Broad, P. Harland, and G. McLean, 2009 Modelling environmental effects on phenology and canopy development of diverse sorghum genotypes. Field Crops Research **111**: 157–165.

Lambert, R., and R. Johnson, 1978 Leaf angle, tassel morphology, and the performance of maize hybrids. Crop Science **18**: 499–502.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, *et al.*, 1987 Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1**: 174–181.

Lee, E., and M. Tollenaar, 2007 Physiological basis of successful breeding strategies for maize grain yield. Crop Science **47**: S–202.

Li, C., Y. Li, Y. Shi, Y. Song, D. Zhang, *et al.*, 2015 Genetic control of the leaf angle and leaf orientation value as revealed by ultra-high density maps in three connected maize populations. PloS one **10**: e0121624.

Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**: 589–595.

Li, P., Y. Wang, Q. Qian, Z. Fu, M. Wang, *et al.*, 2007 *LAZY1* controls rice shoot gravitropism through regulating polar auxin transport. Cell Research **17**: 402–10.

Li, Z., S. R. Pinson, A. H. Paterson, W. D. Park, and J. W. Stansel, 1997 Genetics of hybrid sterility and hybrid breakdown in an intersubspecific rice (*Oryza sativa* L.) population. Genetics **145**: 1139–1148.

Lincoln, S. E., and E. S. Lander, 1992 Systematic detection of errors in genetic linkage data. Genomics **14**: 604–610.

Lister, C., and C. Dean, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. The Plant Journal **4**: 745–750.

Lobell, D. B., G. L. Hammer, K. Chenu, B. Zheng, G. McLean, *et al.*, 2015 The shifting influence of drought and heat stress for crops in northeast australia. Global change biology **21**: 4115–4127.

Lobell, D. B., M. J. Roberts, W. Schlenker, N. Braun, B. B. Little, *et al.*, 2014 Greater sensitivity to drought accompanies maize yield increase in the us midwest. Science **344**: 516–519.

Lobet, G., V. Couvreur, F. Meunier, M. Javaux, and X. Draye, 2014 Plant water uptake in drying soils. Plant Physiology **164**: 1619–1627.

Long, S. P., X. G. Zhu, S. L. Naidu, and D. R. Ort, 2006 Can improvement in photosynthesis increase crop yields? Plant Cell and Environment **29**: 315–330.

Lorieux, M., B. Goffinet, X. Perrier, D. G. de León, and C. Lanaud, 1995a Maximum-likelihood models for mapping genetic markers showing segregation distortion. 1. Backcross populations. Theoretical and Applied Genetics **90**: 73–80.

Lorieux, M., X. Perrier, B. Goffinet, C. Lanaud, and D. G. de León, 1995b Maximum-likelihood models for mapping genetic markers showing segregation distortion. 2. F2 populations. Theoretical and Applied Genetics **90**: 81–9.

Mace, E. S., J.-F. Rami, S. Bouchet, P. E. Klein, R. R. Klein, *et al.*, 2009 A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array Technology (DArT) markers. BMC Plant Biology **9**:

13.

Manichaikul, A., J. Y. Moon, S. Sen, B. S. Yandell, and K. W. Broman, 2009 A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. Genetics **181**: 1077–1086.

Mansfield, B. D., and R. H. Mumm, 2014 Survey of plant density tolerance in us maize germplasm. Crop Science **54**: 157–173.

MATLAB, 2010 *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts.

McAdam, S. A. M., and T. J. Brodribb, 2015 The evolution of mechanisms driving the stomatal response to vapor pressure deficit. Plant Physiology **167**: 833–843.

McAdam, S. A. M., F. C. Sussmilch, T. J. Brodribb, and J. J. Ross, 2015 Molecular characterization of a mutation affecting abscisic acid biosynthesis and consequently stomatal responses to humidity in an agriculturally important species. Aob Plants **7**.

McCormick, R. F., S. K. Truong, and J. E. Mullet, 2015 Rig: recalibration and interrelation of genomic sequence data with the gatk. G3: Genes— Genomes— Genetics **5**: 655–665.

Mccormick, R. F., S. K. Truong, and J. E. Mullet, 2016 3D sorghum reconstructions from depth images identify QTL regulating shoot architecture. Plant Physiology **172**: 823–834.

McCown, R., G. Hammer, J. Hargreaves, D. Holzworth, and D. Freebairn, 1996 APSIM: a novel software system for model development, model testing and simulation in agricultural systems research. Agricultural Systems **50**: 255–271.

McDowell, N. G., A. P. Williams, C. Xu, W. T. Pockman, L. T. Dickman, *et al.*, 2016 Multi-scale predictions of massive conifer mortality due to chronic temperature rise. Nature Climate Change **6**: 295–300.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**: 1297–1303.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li, *et al.*, 2009 Genetic properties of the maize nested association mapping population. Science **325**: 737–740.

Menz, M., R. Klein, J. Mullet, J. Obert, N. Unruh, *et al.*, 2002 A high-density genetic map of *Sorghum bicolor* (L.) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Molecular Biology **48**: 483–499.

Messina, C. D., T. R. Sinclair, G. L. Hammer, D. Curan, J. Thompson, *et al.*, 2015 Limited-transpiration trait may increase maize drought tolerance in the us corn belt. Agronomy Journal **107**: 1978–1986.

Monsi, M., and T. Saeki, 2005 On the factor light in plant communities and its importance for matter production. Annals of Botany **95**: 549–567.

Monteith, J. L., and C. Moss, 1977 Climate and the efficiency of crop production in Britain. Philosophical Transactions of the Royal Society of London B: Biological Sciences **281**: 277–294.

Moon, J., H. Candela, and S. Hake, 2013 The *Liguleless narrow* mutation affects proximal-distal signaling and leaf growth. Development **140**: 405–412.

Moreno, M., L. Harper, R. Krueger, and S. Dellaporta, 1997 *Liguleless1* encodes a nuclear localized protein required for induction of ligules and auricles during maize leaf organogenesis. Genes and Development **11**: 616–628.

Morishige, D. T., P. E. Klein, J. L. Hilley, S. M. E. Sahraeian, A. Sharma, *et al.*, 2013 Digital genotyping of sorghum–a diverse plant species with a large repeat-rich genome. BMC Genomics **14**: 448.

Mravec, J., M. Kubeš, A. Bielach, V. Gaykova, J. Petrášek, *et al.*, 2008 Interaction of

PIN and PGP transport mechanisms in auxin distribution-dependent development. Development **135**: 3345–3354.

Mueller, N. D., J. S. Gerber, M. Johnston, D. K. Ray, N. Ramankutty, *et al.*, 2012 Closing yield gaps through nutrient and water management. Nature **490**: 254–257.

Mullet, J., D. Morishige, R. McCormick, S. Truong, J. Hilley, *et al.*, 2014 Energy sorghuma genetic model for the design of c4 grass bioenergy crops. Journal of Experimental Botany : eru229.

Multani, D. S., S. P. Briggs, M. A. Chamberlin, J. J. Blakeslee, A. S. Murphy, *et al.*, 2003 Loss of an MDR transporter in compact stalks of maize *br2* and sorghum *dw3* mutants. Science **302**: 81–84.

Murchie, E., and M. Reynolds, 2012 *Crop radiation capture and use efficiency*, book section 171. Springer New York, 2615–2638.

Murchie, E. H., Y. Z. Chen, S. Hubbart, S. B. Peng, and P. Horton, 1999 Interactions between senescence and leaf orientation determine *in situ* patterns of photosynthesis and photoinhibition in field-grown rice. Plant Physiology **119**: 553–563.

Murphy, R. L., R. R. Klein, D. T. Morishige, J. A. Brady, W. L. Rooney, *et al.*, 2011 Coincident light and clock regulation of pseudoresponse regulator protein 37 (*PRR37*) controls photoperiodic flowering in sorghum. Proceedings of the National Academy of Sciences of the United States of America **108**: 16469–16474.

Murphy, R. L., D. T. Morishige, J. A. Brady, W. L. Rooney, S. S. Yang, *et al.*, 2014 *Ghd7* (*Ma(6)*) represses sorghum flowering in long days: *Ghd7* alleles enhance biomass accumulation and grain production. Plant Genome **7**.

Nair, S. S., S. J. Kang, X. S. Zhang, F. E. Miguez, R. C. Izaurralde, *et al.*, 2012 Bioenergy crop models: descriptions, data requirements, and future challenges. Global Change Biology Bioenergy **4**: 620–633.

Nobel, P. S., 1999 *Physicochemical and environmental plant physiology*. Academic

press.

Nobel, P. S., 2005 *Physicochemical and environmental plant physiology*. Elsevier Academic Press, Amsterdam ; Boston, 3rd edition.

Olson, S. N., K. Ritter, W. Rooney, A. Kemanian, B. A. McCarl, *et al.*, 2012 High biomass yield energy sorghum: developing a genetic model for c4 grass bioenergy crops. Biofuels, Bioproducts, and Biorefining **6**: 640–655.

Ort, D. R., and S. P. Long, 2014 Limits on yields in the corn belt. Science **344**: 483–484.

Paran, I., I. Goldman, S. Tanksley, and D. Zamir, 1995 Recombinant inbred lines for genetic mapping in tomato. Theoretical and Applied Genetics **90**: 542–548.

Parent, B., C. Hachez, E. Redondo, T. Simonneau, F. Chaumont, *et al.*, 2009 Drought and abscisic acid effects on aquaporin content translate into changes in hydraulic conductivity and leaf growth rate: A trans-scale approach. Plant Physiology **149**: 2000–2012.

Park, S. Y., F. C. Peterson, A. Mosquna, J. Yao, B. F. Volkman, *et al.*, 2015 Agrochemical control of plant water use using engineered abscisic acid receptors. Nature **520**: 545–550.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, *et al.*, 2009 The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**: 551–556.

Pendleton, J., G. Smith, S. Winter, and T. Johnston, 1968 Field investigations of the relationships of leaf angle in corn (zea mays l.) to grain yield and apparent photosynthesis. Agronomy Journal **60**: 422–424.

Peng, B., R. K. Yu, K. L. Dehoff, and C. I. Amos, 2007 Normalizing a large number of quantitative traits using empirical normal quantile transformation. BMC Proc **1 Suppl 1**: S156.

Peng, Y., K. F. Schertz, S. Cartinhour, and G. E. Hart, 1999 Comparative genome mapping of *Sorghum bicolor* (L.) Moench using an RFLP map constructed in a population of recombinant inbred lines. Plant Breeding **118**: 225–235.

Perez, M. B. M., J. Zhao, Y. Yin, J. Hu, and M. G. S. Fernandez, 2014 Association mapping of brassinosteroid candidate genes and plant architecture in a diverse panel of *Sorghum bicolor*. Theoretical and Applied Genetics **127**: 2645–2662.

Porter, J. R., A. Challinor, F. Ewert, P. Falloon, T. Fischer, *et al.*, 2010 Food security: focus on agriculture .

Pradal, C., S. Dufour-Kowalski, F. Boudon, C. Fournier, and C. Godin, 2008 Openalea: a visual programming and component-based software platform for plant modelling. Functional Plant Biology **35**: 751–760.

Prusinkiewicz, P., M. Shirmohammadi, and F. Samavati, 2012 L-systems in geometric modeling. International Journal of Foundations of Computer Science **23**: 133–146.

Purcell, S., and C. Chang, 2014 PLINK. https://www.cog-genomics.org/plink2 .

Riar, M. K., T. R. Sinclair, and P. V. Prasad, 2015 Persistence of limited-transpiration-rate trait in sorghum at high temperature. Environmental and Experimental Botany **115**: 58–62.

Rooney, W. L., and S. Aydin, 1999 Genetic control of a photoperiod-sensitive response in *Sorghum bicolor* (L.) Moench. Crop Science **39**: 397–400.

Rooney, W. L., J. Blumenthal, B. Bean, and J. E. Mullet, 2007 Designing sorghum as a dedicated bioenergy feedstock. Biofuels, Bioproducts and Biorefining **1**: 147–157.

Rosenthal, W., G. Arkin, P. e. Shouse, and W. Jordan, 1987 Water deficit effects on transpiration and leaf growth. Agronomy Journal **79**: 1019–1026.

Sakamoto, T., Y. Morinaka, T. Ohnishi, H. Sunohara, S. Fujioka, *et al.*, 2006 Erect leaves caused by brassinosteroid deficiency increase biomass production and grain yield in rice. Nature Biotechnology **24**: 105–9.

Schmitt, J., J. R. Stinchcombe, M. S. Heschel, and H. Huber, 2003 The adaptive evolution of plasticity: phytochrome-mediated shade avoidance responses. Integrative and Comparative Biology **43**: 459–469.

Shannon, L. M., B. S. Yandell, and K. W. Broman, 2013 Users guide for new BC_sF_t Tools for R/qtl. World Wide Web : 1–20.

Sinclair, T. R., G. L. Hammer, and E. J. Van Oosterom, 2005 Potential yield and water-use efficiency benefits in sorghum from limited maximum transpiration rate. Functional Plant Biology **32**: 945–952.

Sinclair, T. R., and J. E. Sheehy, 1999 Erect leaves and photosynthesis in rice. Science **283**: 1456–1457.

Sinclair, T. R., C. Tanner, and J. Bennett, 1984 Water-use efficiency in crop production. BioScience **34**: 36–40.

Song, Q. F., G. L. Zhang, and X. G. Zhu, 2013 Optimal crop canopy architecture to maximise canopy photosynthetic CO_2 uptake under elevated CO_2 - a theoretical study using a mechanistic model of canopy photosynthesis. Functional Plant Biology **40**: 109–124.

Sybenga, J., 1996 Recombination and chiasmata: few but intriguing discrepancies. Genome **39**: 473–484.

Tange, O., 2011 GNU parallel - the command-line power tool. ;login: The USENIX Magazine **36**: 42–47.

Tardieu, F., 2011 Any trait or trait-related allele can confer drought tolerance: just design the right drought scenario. Journal of Experimental Botany : err269.

Tardieu, F., and W. J. Davies, 1992 Stomatal response to abscisic acid is a function of current plant water status. Plant physiology **98**: 540–545.

Technow, F., C. D. Messina, L. R. Totir, and M. Cooper, 2015 Integrating crop growth models with whole genome prediction through approximate bayesian com-

putation. PloS one **10**: e0130855.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, *et al.*, 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics **43**: 159–62.

Tilman, D., C. Balzer, J. Hill, and B. L. Befort, 2011 Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences **108**: 20260–20264.

Tollenaar, M., and J. Wu, 1999 Yield improvement in temperate maize is attributable to greater stress tolerance. Crop Science **39**: 1597–1604.

Truong, S. K., R. F. McCormick, D. T. Morishige, and J. E. Mullet, 2014 Resolution of genetic map expansion caused by excess heterozygosity in plant recombinant inbred populations. G3: Genes— Genomes— Genetics **4**: 1963–1969.

Truong, S. K., R. F. McCormick, W. L. Rooney, and J. E. Mullet, 2015 Harnessing genetic variation in leaf angle to increase productivity of *Sorghum bicolor*. Genetics **201**: 1229–1238.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, *et al.*, 2013 *From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit Best Practices Pipeline*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Van Oosterom, E., A. Borrell, K. Deifel, and G. Hammer, 2011 Does increased leaf appearance rate enhance adaptation to postanthesis drought stress in sorghum? Crop Science **51**: 2728–2740.

van Zanten, M., T. L. Pons, J. A. M. Janssen, L. A. C. J. Voesenek, and A. J. M. Peeters, 2010 On the relevance and control of leaf angle. Critical Reviews in Plant Sciences **29**: 300–316.

Walsh, J., C. A. Waters, and M. Freeling, 1998 The maize gene *liguleless2* encodes a basic leucine zipper protein involved in the establishment of the leaf blade-sheath

boundary. Genes and Development **12**: 208–218.

Wang, C., C. Zhu, H. Zhai, and J. Wan, 2005 Mapping segregation distortion loci and quantitative trait loci for spikelet sterility in rice (*Oryza sativa* L.). Genetical Research **86**: 97–106.

Warnasooriya, S. N., and T. P. Brutnell, 2014 Enhancing the productivity of grasses under high-density planting by engineering light responses: from model systems to feedstocks. Journal of Experimental Botany **65**: 2825–2834.

Wu, R., C. Ma, and G. Casella, 2007 *Statistical genetics of quantitative traits: Linkage, maps and QTL*. Springer.

Xin, Z., D. Gitz, G. Burow, C. Hayes, and J. Burke, 2015 Registration of two allelic erect leaf mutants of sorghum. Journal of Plant Registrations **9**: 254–257.

Xu, S., 2003 Theoretical basis of the beavis effect. Genetics **165**: 2259–2268.

Xu, S., 2008 Quantitative trait locus mapping can benefit from segregation distortion. Genetics **180**: 2201–2208.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2013 Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Genome-wide association studies and genomic prediction : 215–236.

Yang, S., B. D. Weers, D. T. Morishige, and J. E. Mullet, 2014 CONSTANS is a photoperiod regulated activator of flowering in sorghum. BMC Plant Biology **14**: 148.

Yoshida, S., 1972 Physiological aspects of grain yield. Annual Review of Plant Physiology **23**: 437–445.

Zhang, J., L. X. Ku, Z. P. Han, S. L. Guo, H. J. Liu, *et al.*, 2014 The *ZmCLA4*

gene in the qLA4-1 QTL controls leaf angle in maize (*Zea mays* L.). Journal of Experimental Botany **65**: 5063–76.

Zhao, S. Q., J. A. Hu, L. B. Guo, Q. A. Qian, and H. W. Xue, 2010 *Rice leaf inclination2*, a VIN3-like protein, regulates leaf angle through modulating cell division of the collar. Cell Research **20**: 935–947.

Zhu, C., C. Wang, and Y.-M. Zhang, 2007 Modeling segregation distortion for viability selection I. reconstruction of linkage maps with distorted markers. Theoretical and Applied Genetics **114**: 295–305.

Zhu, X.-G., S. P. Long, and D. R. Ort, 2008 What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? Current Opinion in Biotechnology **19**: 153–159.

Zhu, X. G., S. P. Long, and D. R. Ort, 2010 Improving photosynthetic efficiency for greater yield. Annual Review Plant Biology **61**: 235–61.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR RESOLUTION OF GENETIC MAP EXPANSION CAUSED BY EXCESS HETEROZYGOSITY IN PLANT RECOMBINANT INBRED POPULATIONS [1]

Plant	**Generation**	**% Observed**	**% Expected**	**Citation**
Arabidopsis	F_8	0.42	0.78125	Lister and Dean (1993)
Tomato	F_7	15.0	1.5625	Paran *et al.* (1995)
Maize	F_{10}	1.6	0.1953125	Burr and Burr (1991)
Maize	F_{10}	2.7	0.1953125	Burr and Burr (1991)
Sorghum	F_6	4.69	3.125	Peng *et al.* (1999)
Sorghum	F_5	19.76*	6.25	Kong *et al.* (2013)

Table A.1: **Reports of deviation from expected heterozygosity maintained per generation.** The 19.76% from Kong et al. comes from an average of the reported distorted regions across the genome; the total proportion of heterozygosity could not be located in the publication, so this is likely an overestimate.

A.1 Solving for the general solution, p_{F_t}, in MATLAB

Given the theory derived for ${p_{F_t}}' = \mathbf{T}{p_{F_{t-1}}}'$ we solved for the general solution of p_{F_t} using MATLAB (2010) and an M-file is provided as a supplemental file to document all variables defined and calculations. The M-file can be found on `https://github.com/MulletLab/exHet_Supplement`, and we also provide it below.

---

[1] Reprinted from S. K. Truong, McCormick, R. F., Morishige, D.M., and J. E. Mullet, 2014. Resolution of Genetic Map Expansion Caused by Excess Heterozygosity in Plant Recombinant Inbred Populations. G3: Genes — Genomes — Genetics 4: 1963-1969 under the Creative Commons Attribution Unported License (`http://creativecommons.org/licenses/by/3.0`), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Copyright ©2014 Truong & McCormick *et al.*

Marker #	**Reported map size (cM)**	**Citation**	**Marker Type**
145	1279	Hart *et al.* (2001)	RFLP, SSR
323	1347	Peng *et al.* (1999)	RFLP
466	1406	Bhattramakki *et al.* (2000)	RFLP, SSR
792	1528	Mace *et al.* (2009)	DaRT, RFLP, SSR
2926	1713	Menz *et al.* (2002)	AFLP, RFLP, SSR

Table A.2: **Reported genetic map sizes for the sorghum BTx623 x IS3620c RIL population used in this study.**

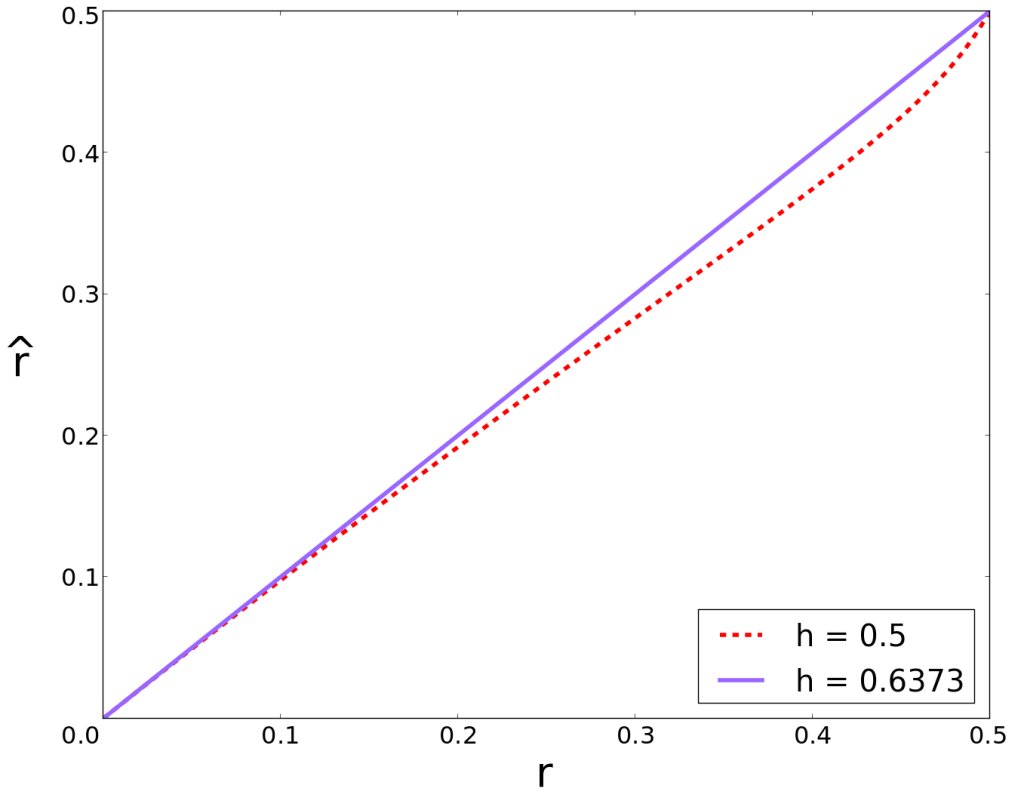

Figure A.1: **Estimated recombination fractions, $\hat{r}$, of excess heterozygosity versus Mendelian expectations for $t = 3$.** Recombination fractions estimated from genotype frequencies under Mendelian expectations (h=0.5) versus under modeling a global heterozygosity advantage (h=0.6373) at generation $t = 3$ of a selfing population. This shows that if the population was retaining excess heterozygosity (at a rate of 63.73% each generation as opposed to the Mendelian 50%), then estimating recombination fractions under Mendelian expectations would shrink the map if observed at generation $t = 3$.

% MATLAB M-file to derive the general solution for the probability of

% a marker in a genotype class given a selfing population, F_t, for t

% generations. This is supplemental information for Truong & McCormick

% et al (2014) where we incorporate a heterozygosity zygotic viability term.

```
%
    % assign variables:
% r is the recombination frequency, and t is the generation interval
 syms r t;
    % h is amount of heterozygosity maintained in each generation and can be
% parameterized given generation t. That is if H is the amount of
% heterozygosity in an F_t population, then h^(t-1)=H
 syms h;
    % u is the viability of Aa to AA and aa
% solve(h == ((2*u^2)*((1-r)^2+r^2) + (2*2*u*r*(1-r)))/d , u )
 syms u;
u = -(2*h*r - r + ((r^2 - 2*h*r + h)*(2*h*r - 2*r - h + r^2 + 1))^(1/2) - 2*h*r^2
+ r^2)/(h + 2*r - 2*h*r + 2*h*r^2 - 2*r^2 - 1);
    % d is a parameter necessary to weigh to u appropriately
 syms d;
d = 2*((1-r)^2)+8*u*r*(1-r)+ 2*(r^2)+ 2*(u^2)*(((1-r)^2)+(r^2);
% Transition probability matrix for 5 classes of genotypes
 T = [
1, 0, (1-h^1)/2, (2*((1-r^1)^2))/d^1, (2*(r^2))/d^1;
0, 1, (1-h^1)/2, (2*(r^2))/d^1, (2*((1-r^1)^2))/d^1;
0, 0, (h^1), (8*u^1*r^1*(1-r^1)^1)/d^1, (8*u^1*r^1*(1-r^1)^1)/d^1;
0, 0, 0, (2*u^2*(1-r^1)^2)/d^1, (2*u^2*r^2)/d^1;
0, 0, 0, (2*u^2*r^2)/d^1, (2*u^2*(1-r^1)^2)/d^1];
% Take eigenvalues of Transition probability matrix to set up system of
% equations to find the general solution given generation t for all 5
% classes
```

```
eigT = eig(T);
% qit=[
% p(class 1 in generation t);
% p(class 2 in generation t);
% p(class 3 in generation t);
% p(class 4 in generation t);
% p(class 5 in generation t)];
% Initialize probability of class given generation t. For an F_1 (t=1)
% from the initial mating of homozygous parents (ie AABB x aabb), all
% individuals in the F_1 are of class 4 (ie AaBb in coupling (AB/ab))
qi1=[0;0;0;1;0];
qi2 = T*qi1;
qi3 = T*qi2;
qi4 = T*qi3;
% bclass = [
% p(class in F1);
% p(class in F2);
% p(class in F3);
% p(class in F4)];
% Set up the frequences directly in F_1, F_2, F_3, and F_4 for each
% class
b1=[qi1(1,1);qi2(1,1);qi3(1,1);qi4(1,1)];
b2=[qi1(2,1);qi2(2,1);qi3(2,1);qi4(2,1)];
b3=[qi1(3,1);qi2(3,1);qi3(3,1);qi4(3,1)];
b4=[qi1(4,1);qi2(4,1);qi3(4,1);qi4(4,1)];
b5=[qi1(5,1);qi2(5,1);qi3(5,1);qi4(5,1)];
```

```
% Set up the 4 linear equations (for each generation t=1,2,3,4)
 A=[
eigT(1,1)^1 eigT(2,1)^1 eigT(3,1)^1 eigT(4,1)^1;
eigT(1,1)^2 eigT(2,1)^2 eigT(3,1)^2 eigT(4,1)^2;
eigT(1,1)^3 eigT(2,1)^3 eigT(3,1)^3 eigT(4,1)^3;
eigT(1,1)^4 eigT(2,1)^4 eigT(3,1)^4 eigT(4,1)^4];
% We now have a system of 4 linear equations with 4 unknowns for each class
% A*[coefficients of general solution]=bclass
 x1=linsolve(A,b1);
x2=linsolve(A,b2);
x3=linsolve(A,b3);
x4=linsolve(A,b4);
x5=linsolve(A,b5);
q_it=[eigT(1,1)^t eigT(2,1)^t eigT(3,1)^t eigT(4,1)^t];
% pclass is the probability of class i (where i=1,2,3,4,5) given
% heterozygosity maintained h, recombination r, and generation t
 p1=q_it*x1;
p2=q_it*x2;
p3=q_it*x3;
p4=q_it*x4;
p5=q_it*x5;
```

A.2 Parameterization of the heterozygosity term

To generate the sorghum genetic map presented in this study, we modeled a global heterozygosity maintained per generation parameter, h, based on the average

heterozygosity observed, H. We briefly discussed the possibility of more local estimations of h, and here we explore the topic in greater detail. Here, we (i) analyze the distribution of heterozygosity, (ii) provide further reasoning to using a global heterozygosity term as well as (iii) present two methods for parameterizing more local fluctuations in h. The first method estimates an h for each linkage group and is implemented as an option in est.rf.exHet(), and the second method derives an h for each marker pair. We also acknowledge an intermediate approach whereby local heterozygosity could be estimated on a regional basis to parameterize a regional h, perhaps with a sliding window.

A.2.1 A global heterozygosity term

Figures A.1 and A.2.1 show that there are regions of variable heterozygosity such that groups of markers vary in their proportion of individuals heterozygous relative to the genome-wide average. When modeling recombination fractions under either Mendelian or excess heterozygosity (as done in the paper) it is assumed that genotypes are uniformly distributed. However, we find that the proportion of heterozygous individuals at a marker more closely follows a normal distribution, suggesting that the assumption of a uniform distribution underlying both models may need to be revisited (Figure A.2.1). However, given (i) that the proportions of heterozygosity are greater than those expected under the Mendelian model genome-wide (Figures A.1 and A.2.1) and (ii) the precedence for assuming a uniform distribution used when estimating recombination fractions under the Mendelian model, we found the use of a global heterozygosity parameter taken from the average of all markers' genotypes to be a reasonable choice.

A.2.2 Local heterozygosity term for each linkage group, $h_{linkage\ group\ \#}$

An alternative to map estimation using a global heterozygosity term is to estimate h for each linkage group. We implemented this alternative, and in our use cases, employing a local parameterization of the heterozygosity based on linkage groups gave similar results to the global heterozygosity parameterization (see Figure A.3 and spreadsheet provided at `https://github.com/MulletLab/exHet_Supplement`). Calculating recombination fractions with local heterozygosity based on linkage groups can be used by invoking est.rf.exHet(hetByLinkageGroup=TRUE). An example of how to call the function used to parameterize h by linkage groups is provided in the example code at `https://github.com/MulletLab/exHet_Supplement`.

A.2.3 The derivation for a local heterozygosity term for each marker, h_{marker}

Here we briefly discuss a general solution, $p_{F_t}{}' = \mathbf{T}p_{F_{t-1}}{}'$, to be solved for in order to incorporate differential heterozygosity for each marker. While we derive it here, we chose not to use it and did not implement it due to the pitfalls associated with overfitting data.

This follows the theory described in the paper such that we will build the transition probability matrix and then solve for the general solution of p_{F_t}. First we will redefine our genotype classes. We are going to treat different markers with differential heterozygosity terms, so it would be nice to split the single heterozygote class (enumerated class 3 in the paper) to class 3_α and class 3_β. Such that now we have

$$p_{F_t} = \begin{pmatrix} p(\text{class 1}) \\ p(\text{class 2}) \\ p(\text{class 3}_\alpha) \\ p(\text{class 3}_\beta) \\ p(\text{class 4}) \\ p(\text{class 5}) \end{pmatrix}_t = \begin{matrix} p(\text{genotypes}) \\ \begin{pmatrix} p(\frac{\text{AB}}{\text{AB}}) + p(\frac{\text{ab}}{\text{ab}}) \\ p(\frac{\text{Ab}}{\text{Ab}}) + p(\frac{\text{aB}}{\text{aB}}) \\ p(\frac{\text{AB}}{\text{aB}}) + p(\frac{\text{Ab}}{\text{ab}}) \\ p(\frac{\text{AB}}{\text{Ab}}) + p(\frac{\text{aB}}{\text{ab}}) \\ p(\frac{\text{AB}}{\text{ab}}) \\ p(\frac{\text{Ab}}{\text{aB}}) \end{pmatrix}_t \end{matrix}$$

A.2.3.1 class 1 and class 2

The transition from class 1 and class 2 in generation t to generation $t+1$ are fixed.

A.2.3.2 class 3

The transition from class 3 in generation t to generation $t+1$ will take into consideration only the segregation of one marker that is heterozygote in generation t as the other marker will be homozygote and thus fixed in any subsequent generation after t.

If H_{α,F_t} proportion of heterozygosity observed in marker α for an F_t family and we assume that the amount of heterozygosity maintained in marker α, h_α, each generation prior to generation t is the same, then we can solve for h_α through the following relationship ${h_\alpha}^{t-1} = H_{\alpha,F_t}$. h_α will be modeled into the transition probability matrix as a modifier of expected segregation. To do so, we can treat marker α's genotypes (zygotes) with differential viability (expectation to be observed in the next generation). Define the amount of heterozygosity maintained at marker α as h_α (parameterized from data as shown above) through selfing. Then our expected

segregation ratio for $AA : Aa : aa$ is

$$\frac{1-h_\alpha}{2} : h_\alpha : \frac{1-h_\alpha}{2}.$$

Notice that under the assumption of Mendelian segregation, $h_\alpha = 1/2$ and the expected Mendelian segregation would then be the familiar $1:2:1$. The same model is true for marker β.

A.2.3.3 class 4 and 5

The transition from class 4 and 5 in generation t to generation $t+1$ will take into consideration both the segregation of two markers that are heterozygous at generation t and the recombination frequency between the two markers.

Similar to treatment of heterozygosity for one marker, we now have a heterozygosity term for both marker α, h_α, and marker β, h_β. Given two heterozygosity terms (one for each marker), we can parameterize both h_{marker}'s for each pair of markers in genetic map construction. Now, in the context of zygotic differential viability, assume that

1. the viability of genotype Aa relative to AA or aa is u_α (dependent on h_α)
2. the viability of genotype Bb relative to BB or bb is u_β (dependent on h_β)

such that the expected segregation is now

$$\mathbf{prob}(\text{genotype }_t \,|\, \text{class 4 }_{t-1}) = \begin{array}{c} \\ \text{BB} \\ \text{Bb} \\ \text{bb} \end{array} \begin{array}{c} \begin{array}{ccc} \text{AA} & \text{Aa} & \text{aa} \end{array} \\ \begin{pmatrix} \frac{1}{d}(1-r)^2 & \frac{2u_\alpha}{d} r(1-r) & \frac{1}{d} r^2 \\ \frac{2u_\beta}{d} r(1-r) & \frac{2u_\alpha u_\beta}{d}[r^2 + (1-r)^2)] & \frac{2u_\beta}{d} r(1-r) \\ \frac{1}{d} r^2 & \frac{2u_\alpha}{d} r(1-r) & \frac{1}{d}(1-r)^2 \end{pmatrix} \end{array}$$

where $d = 2(1-r)^2 + 4u_\alpha r(1-r) + 4u_\beta r(1-r) + 2r^2 + 2u_\alpha u_\beta[(1-r)^2 + r^2]$. Then, the amount of heterozygosity retained in generation t for a marker pair of either class 4 or 5[2] in the previous generation $t-1$ should satisfy

$$h = \frac{1}{2}\text{prob(AaBB)} + \frac{1}{2}\text{prob(AABb)} + \text{prob(AaBb)} + \frac{1}{2}\text{prob(aaBb)} + \frac{1}{2}\text{prob(Aabb)}$$

such that given data H_{α,F_t}, H_{β,F_t}, and t we can calculate

$$h_{\text{marker}} = e^{\frac{\ln(H_{\text{marker}},F_t)}{t-1}}.$$

Furthermore, given data r for each marker pair we can subsequently calculate u_α, u_β and d .

A.2.3.4 Transition probability matrix

Incorporating the transition from a class # to other classes (from the previous sections) in every generation, we now have a transition probability matrix,

$$\mathbf{T} = \begin{array}{c} \\ \text{class 1} \\ \text{class 2} \\ \text{class } 3_\alpha \\ \text{class } 3_\beta \\ \text{class 4} \\ \text{class 5} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{class 1} & \text{class 2} & \text{class } 3_\alpha & \text{class } 3_\beta & \text{class 4} & \text{class 5} \end{array} \\ \begin{pmatrix} 1 & 0 & \frac{1-h_\alpha}{2} & \frac{1-h_\beta}{2} & \frac{2(1-r)^2}{d} & \frac{2r^2}{d} \\ 0 & 1 & \frac{1-h_\alpha}{2} & \frac{1-h_\beta}{2} & \frac{2r^2}{d} & \frac{2(1-r)^2}{d} \\ 0 & 0 & h_\alpha & 0 & \frac{4u_\alpha r(1-r)}{d} & \frac{4u_\alpha r(1-r)}{d} \\ 0 & 0 & 0 & h_\beta & \frac{4u_\beta r(1-r)}{d} & \frac{4u_\beta r(1-r)}{d} \\ 0 & 0 & 0 & 0 & \frac{2u_\alpha u_\beta(1-r)^2}{d} & \frac{2u_\alpha u_\beta r^2}{d} \\ 0 & 0 & 0 & 0 & \frac{2u_\alpha u_\beta r^2}{d} & \frac{2u_\alpha u_\beta(1-r)^2}{d} \end{pmatrix} \end{array}$$

With $\mathbf{T}$ defined, we can solve for the general solution of p_{F_t} dependent on t, r,

[2] A probability matrix of all genotypes created after selfing of class 5 would look similar except for the exchange of rows 1 and 3.

and H_{F_t} for every marker pair. Solving for the general solution here is conceptually similar to the process described for the global heterozygosity term and described in the M-file in section 1.

A.3 Genetic map estimations of simulated datasets with excess heterozygosity

This section describes a simulation study performed to demonstrate the effect of accounting for excess heterozygosity in the genetic model. Figure A.3 and the associated spreadsheet found at `https://github.com/MulletLab/exHet_Supplement` provide the results of estimating genetic maps for the sorghum mapping population and the simulated data using different models and methods. The following factors were considered:

1. Dataset
    (a) BTx623 x IS3620c with tight double recombinations removed
    (b) BTx623 x IS3620c without tight double recombinations removed
    (c) Simulated data generated under conditions of excess heterozygosity
    (d) Simulated data generated under conditions of excess heterozygosity with 1% error rate and 5% missing data
    (e) Simulated data generated under conditions of excess heterozygosity with 1% error rate and 5% missing data with tight double recombinations removed
2. Method
    (a) Pairwise estimation using est.rf() or est.rf.exHet(), where est.rf.exHet(h = 0.5) is equal to est.rf()

(b) Multipoint estimation with a hidden Markov model using est.map() and a 1% error probability

3. Model

   (a) Mendelian model ($h = 0.5$)

   (b) Derived heterozygosity model, global h ($h = 0.6373$)

   (c) Derive heterozygosity model, local h by linkage group (hetByLinkageGroup=TRUE)

4. Generation Interval

   (a) F_7

   (b) Fixed RIL ($t \to \infty$)

A.3.1 Tight double recombinations

Tight double recombinations, also referred to as short double crossovers (SDCOs) in the provided code and results, are most often treated as genotyping errors. In the sorghum mapping population dataset used in this paper, setting short double crossovers smaller than 2 cM to missing removed 1.1% of the genotypes (37,299 out of 3,407,539). When simulating genotyping errors, we used a 1% error rate. The method we used to remove short double crossovers was sufficient to compensate for a 1% error rate in the simulated dataset, such that both pairwise estimation (i.e. using est.rf.exHet()) and multipoint estimation using the HMM (i.e. using est.map()) provided comparable results between (i) the simulated data with 1% error rate with tight double recombinations removed and (ii) the simulated data without error. The method we used to remove tight double recombinations is provided as a Python script at `https://github.com/MulletLab/exHet_Supplement`.

A.3.2 Error probability in the hidden Markov model

The results from map estimation using a hidden Markov model (HMM) as implemented in R/qtl's est.map() function on the simulated datasets and the sorghum mapping population showed that the HMM methodology handled the error rate effectively in the simulated data, compensating for the 1% error rate in the simulated data and yielding very similar results whether or not tight double recombinations were removed (though still giving expanded maps since the underlying genetic model assumed $h = 0.5$). However, the HMM multipoint methodology yields very different results between the sorghum mapping dataset with and without tight double recombinations removed. Both the multipoint and the pairwise give grossly inflated maps if tight double recombinations are not removed ($>$ 3000 cM), and give comparable results once tight double crossovers are removed (around 1600 cM). This suggests that the random errors introduced in the simulation were not representative of the errors in sorghum mapping dataset.

A.3.3 F7 versus fixed RIL ($t \to \infty$)

Removal of all heterozygous genotypes (e.g. treating the map as a fixed RIL) reduces the map size for both the simulated and real datasets (under both pairwise and multipoint methods). This is expected since the removal of heterozygous genotypes effectively removes recombination events; in the case of our real dataset this omits 6.7% of the genotypes, or 224,437 genotype calls. For the simulated dataset, the estimated map is still larger than the simulated linkage group since the underlying genetic model does not account for the excess heterozygosity. Unlike the simulated dataset, the multipoint method (i.e. est.map()) yields very different results between treating the real data as an F7 and as a fixed RIL, especially if tight double recombinants are not removed. We suspect that this may be a consequence of the error

modeled by the HMM not being representative of how the error exists in the read data.

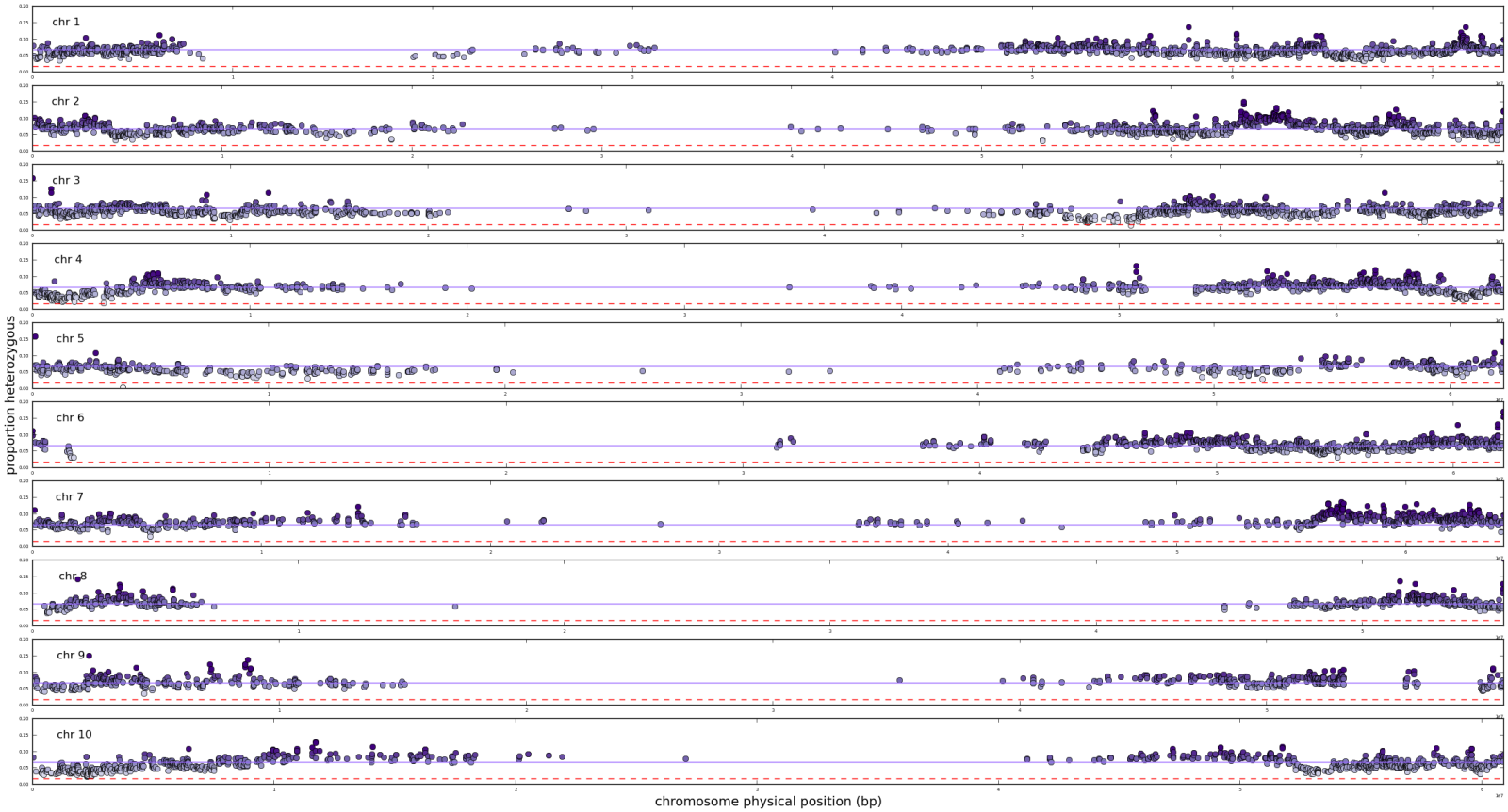

Figure A.2: **Heterozygosity landscape.** Dot plot of the proportion of heterozygous genotypes versus the physical base pair position of the 10,081 markers. The coloring of the markers correspond to the percentage of heterozygosity as explained in Figure 2.3. The Mendelian expected proportion of heterozygosity of an F_7 RIL population is 0.016 and the observed heterozygosity as an average of the BTx623×IS3620c F_7 is 0.067 depicted by a red dashed line and purple solid line, respectively.

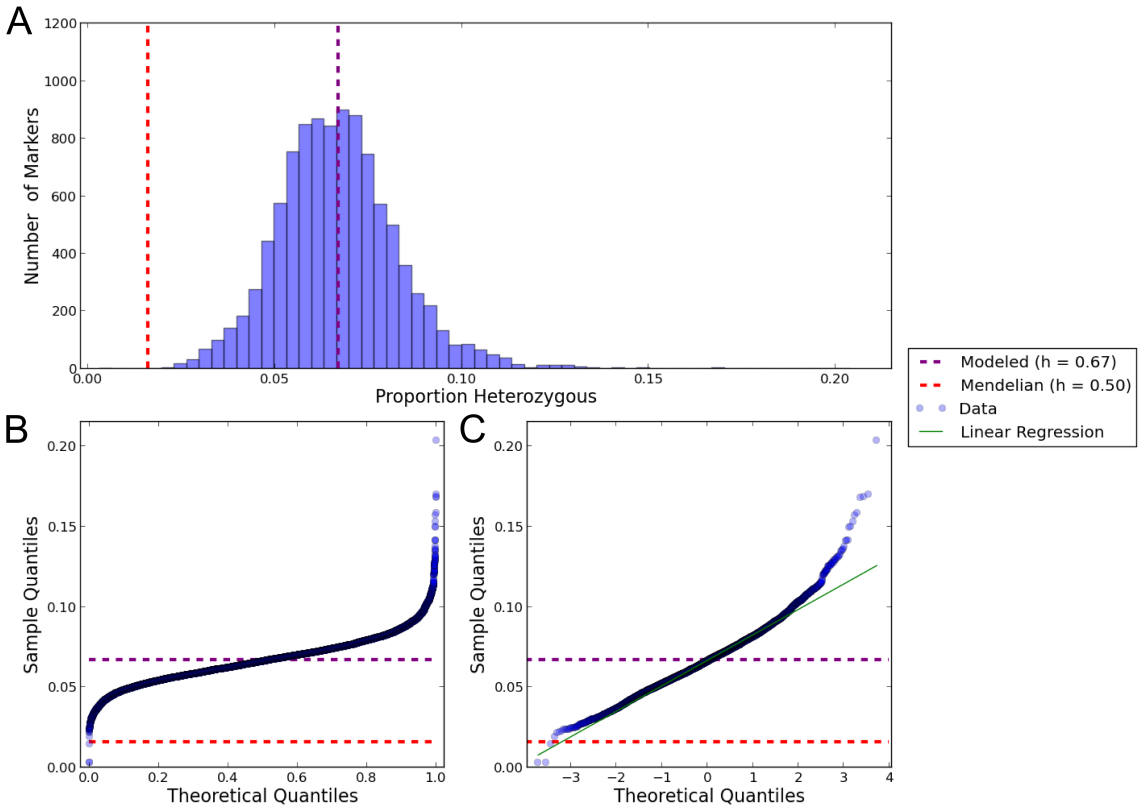

Figure A.3: **Heterozygosity distribution in sorghum F7 mapping population.** (**A**) A rough assessment of the distribution of (excess) heterozygosity of the markers used to genotype the sorghum mapping population shows that most markers display more heterozygosity than expected under Mendelian assumptions of segregation (depicted by a red dashed line). The histogram also shows the average excess heterozygosity (depicted by a purple dashed line) that was used to estimate recombination in the RIL. Quantile-Quantile (Q-Q) plots compare the heterozygosity distributions against (**B**) a uniform distribution and (**C**) a normal distribution. By plotting sample quantiles against theoretical quantiles for the distributions, it can be argued that the excess heterozygosity appears to be more normally distributed than it is uniformly distributed.

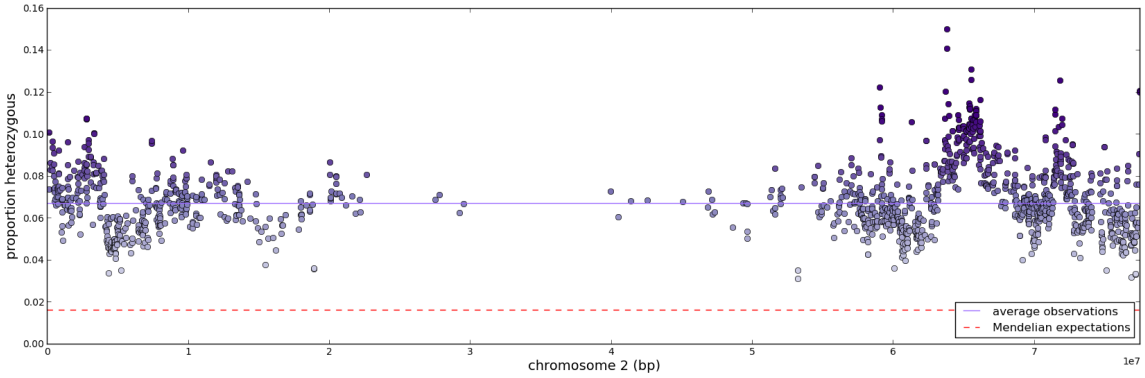

Figure A.4: **Heterozygosity on chromosome 2.** Dot plot of the proportion of heterozygous genotypes vs the physical base pair position of markers on chromosome 2 illustrates the variability of heterozygosity observed in regions of the genome.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	BTx623 x IS3620c – short double crossovers (SDCO) set to missing																	
2						F7 est.rf.exHet(h=0.6373)		F7 est.rf.exHet(hetByLinkageGroup=TRUE)			F7 est.rf.exHet(h=0.5)		F7 est.rf()		F7 est.map(errorprob=0.01)		RIL est.map(errorprob=0.01)	
3	Linkage Group	Marker #				cM distance		estimated h		cm distance		cM distance		cM distance		cM distance		cM distance	
4		1	1418			177		0.6350		177.6		206.7		206.7		203.0		146.5	
5		2	1398			185.9		0.6399		185.3		213.4		213.4		211.4		151.5	
6		3	1495			179.2		0.6234		182.6		208.6		208.6		207.2		143.9	
7		4	1039			146.2		0.6350		146.6		169.2		169.2		172.4		128.3	
8		5	663			109.4		0.6273		110.8		126.2		126.2		132.8		98.1	
9		6	1007			117.6		0.6419		116.8		135.7		135.7		145.5		105.5	
10		7	804			113		0.6551		110.6		127.5		127.5		133.5		103.3	
11		8	528			101		0.6451		100.0		114.7		114.7		122.6		90.3	
12		9	641			118.7		0.6441		117.6		135.3		135.3		138.1		101.9	
13		10	1088			142.8		0.6291		144.4		166.6		166.6		162.9		107.4	
14	Total		10081			1390.6				1392.3		1603.8		1603.8		1629.3		1176.7	
15																			
16	BTx623 x IS3620c - with short double crossovers (SDCO)																		
17						F7 est.rf.exHet(h=0.6373)		F7 est.rf.exHet(hetByLinkageGroup=TRUE)			F7 est.rf.exHet(h=0.5)		F7 est.rf()		F7 est.map(errorprob=0.01)		RIL est.map(errorprob=0.01)	
18	Linkage Group	Marker #				cM distance		estimated h		cm distance		cM distance		cM distance		cM distance		cM distance	
19		1	1418			468.7		0.6467		463.1		534.7		534.7		452.9		160.1273	
20		2	1398			461.4		0.6522		453.1		521.6		521.6		435.5		158.2769	
21		3	1495			498.2		0.6381		497.7		571.3		571.3		481.7		151.9861	
22		4	1039			370.8		0.6476		366.1		421.6		421.6		362.7		134.1605	
23		5	663			241.4		0.6407		240.3		275.8		275.8		247.4		105.2321	
24		6	1007			322.5		0.6544		315.8		364.8		364.8		312.1		113.5757	
25		7	804			269.2		0.6650		260.9		299.2		299.2		258.8		110.3455	
26		8	528			205.5		0.6559		201.0		230.5		230.5		217.4		95.44168	
27		9	641			251.2		0.6556		245.8		281.7		281.7		251.3		107.4022	
28		10	1088			371.9		0.6426		369.3		426.4		426.4		365.3		111.5893	
29	Total		10081			3460.9				3413.2		3927.6		3927.6		3385.2		1248.137	
30																			
31	Simulated (n=1000, t=7, h=0.6373) – No errors, no missing data																		
32				Simulated Distance		F7 est.rf.exHet(h=0.6373)		F7 est.rf.exHet(hetByLinkageGroup=TRUE)			F7 est.rf.exHet(h=0.5)		F7 est.rf()		F7 est.map(errorprob=0.01)		RIL est.map(errorprob=0.01)	
33	Linkage Group	Marker #		cM distance		cM distance		estimated h		cm distance		cM distance		cM distance		cM distance		cM distance	
34		1	1000	200		204.9		0.6346		205.6		236.0		236.0		234.0		227.8	
35																			
36	Simulated (n=1000, t=7, h=0.6373) – Errors (1%) and missing data (5%) (i.e. With SDCOs)																		
37				Simulated Distance		F7 est.rf.exHet(h=0.6373)		F7 est.rf.exHet(hetByLinkageGroup=TRUE)			F7 est.rf.exHet(h=0.5)		F7 est.rf()		F7 est.map(errorprob=0.01)		RIL est.map(errorprob=0.01)	
38	Linkage Group	Marker #		cM distance		cM distance		estimated h		cm distance		cM distance		cM distance		cM distance		cM distance	
39		1	1000	200		742.8		0.6375		742.6		854.2		854.2		240.2		237.6	
40																			
41	Simulated (n=1000, t=7, h=0.6373) – Errors (1%) and Missing Data (5%) - SDCOs set to missing																		
42				Simulated Distance		F7 est.rf.exHet(h=0.6373)		F7 est.rf.exHet(hetByLinkageGroup=TRUE)			F7 est.rf.exHet(h=0.5)		F7 est.rf()		F7 est.map(errorprob=0.01)		RIL est.map(errorprob=0.01)	
43	Linkage Group	Marker #		cM distance		cM distance		estimated h		cm distance		cM distance		cM distance		cM distance		cM distance	
44		1	1000	200		196.3		0.6340		197.1		226.3		226.3		231.6		224.5	

Figure A.5: **Screenshot of spreadsheet containing map estimation results for the sorghum mapping population and simulated data with different models and methods.** This spreadsheet is provided as a .ods and a .xlsx file at `https://github.com/MulletLab/exHet_Supplement` and a description of its results are here in the Supplemental Information text.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR HARNESSING GENETIC VARIATION IN LEAF ANGLE TO INCREASE PRODUCTIVITY OF *SORGHUM BICOLOR* [1]

B.1 Extended materials and methods

B.1.1 Calculation of the light extinction coefficient, k.

Light interception throughout crop canopies is often formalized as an extinction coefficient, k (as derived in Beer-Lambert's Law), that relates the attenuation of light with properties of the material through which the light travels (Monteith and Moss, 1977; Monsi and Saeki, 2005; Nobel, 1999; Long *et al.*, 2006). Here, we describe adoption of Beer-Lambert equations in the context of solar radiation attenuation through a sorghum canopy to characterize and compare the distribution of light in simulated and field grown sorghum canopies with differing leaf inclination angles.

Let $I(x)$ be the intensity, power per unit area, of radiation from the sunlight at depth x down the canopy. For any given time at a plane above the canopy, $x = 0$, the intensity of radiation from the sun is at its maximum as no radiation has been intercepted, $I(0) = \max_x I(x)$. A sequence of planes, parallel to the ground, moving from the top, $x = 0$, down to the ground level of the canopy will have decreasing amounts of transmitted radiation, as each plane in the canopy intercepts a portion of the radiation. As such, the rate of radiation or light extinction when moving down

[1] Reprinted with permission from "Harnessing genetic variation in leaf angle to increase productivity of *Sorghum bicolor*" by Truong, SK, McCormick, RF, Rooney, WL, and Mullet, JE, (2015) Genetics 201(3) 1229-1238, Copyright ©2015 by the Genetics Society of America.

the canopy, $\frac{dI}{dx}$, can be represented as

$$[h!]h = \frac{dI}{dx} = kI(x) \tag{B.1}$$

where k is a dimensionless variable that represents the proportion of radiation intercepted at depth x down the canopy. By integration of Equation 2.1 and the initial condition $I(0) = I_0$,

$$[h!]I(x) = I_0 \exp -kx \tag{B.2}$$

Thus given data, $x_n, I(x_n)_{n=1}^{N}$, that can be reasonably described by Equation B.2, we can find light extinction coefficient, k, and use this as a descriptor for the distribution of light down a sorghum canopy. We defined I_0, the top of canopy, to be the plane immediately below the whorls of plants in the canopy. We distinguish between two groups of leaves in a sorghum canopy: (i) the whorl, the top of the plant from which leaves emerge, (ii) and leaves that are fully expanded below the whorl. The whorl contains leaves that are essentially in the same plane, whereas fully expanded leaves share less planes with other leaves above and below it. We found better fits to light data when fitting data just below the whorl to find the light extinction coefficient, k, however the relative relationship between small and large leaf inclinations angles remained regardless of where we defined the top of the canopy. The next sections will describe the method used to collect the data in both simulated and field experiments to estimate light extinction coefficients, ks. Specifically, k values for simulated and experimental data were solved for by the Levenberg-Marquardt algorithm for non-linear least-squares as implemented in the open source software SciPy's scipy.optimize.curve_fit() function (Jones 2001).

B.1.2 Virtual sorghum canopies.

In order to calculate theoretical ks we constructed functional-structural plant models of sorghum and collected depth and incident light data, $x_n, I(x_n)_{n=1}^{N}$, in simulated light environments. The 3-dimensional virtual sorghum plants were constructed using Lindenmeyer systems in L-py (Boudon *et al.*, 2012). Lindenmeyer-systems provide a set of production rules whereby plant structural models are produced by recursion through phytomers (Prusinkiewicz *et al.*, 2012). As such, virtual sorghum was constructed from phytomers, characterized by a stem and a leaf that emerges on top of one other. Thus each sorghum plant has a set of phytomers $p = p_1, p_2, ..., p_N$ where p_1 is the first phytomer to emerge and closest to the bottom and p_N is the last fully emerged phytomer at the top of the plant; only the colored phytomers are considered to calculate k (Figure 2a). These virtual plants were then replicated to simulate sorghum plots that have typical row spacing of 0.76 m and planting density of 13.2 plants/m2. Since the phytomers appear one on top of another and are uniformly distanced from the ground in the virtual environment, fully emerged phytomers, p_i, are used as depth measurements in the virtual sorghum canopies. When the simulated sorghum genotypes varied in the number of the phytomers (and consequently height) as they did when simulating plants with the characteristics of RIL 63 and RIL 73 of the R07018 x R07020 recombinant inbred line population, the height difference was accounted for by scaling depth to be a percentage depth, such that xn=100pnpN . The qualitative relationship of the rates of extinction between canopies with large and small leaf angles remains the same whether or not scaling is performed; scaling height just removes an unnecessary complexity.

The virtual canopies were then illuminated by the nested radiosity model as implemented in CARIBU (Chelle and Andrieu, 1998; Chelle *et al.*, 2004) given light

input that reflected solar conditions in College Station, TX (data retrieved from The United States Naval Observatory). The illuminated sorghum canopies could then be visualized (see Supporting Information for a .gif time course of lighting over a day). The nested radiosity model enables the estimation of absorbed energy or irradiance of each phytomer. Construction and lighting of the virtual canopies were all done within the open source software OpenAlea v1.0 that integrates L-py and Caribu, among other functional-structural modeling tools (Pradal *et al.*, 2008). Let $E_{abs}(p_n)$ be the energy absorbed by phytomer n, p_n. Then to extract the amount of energy or incident light available, $I(x_n)$, at x_n

$$[h!]I(x_n) = \sum_{i=1}^{N} E_{abs}(p_i) - \sum_{i=n}^{N} E_{abs}(p_i). \tag{B.3}$$

With Equation B.2 and pn defined, this provides data, $x_n, I(x_n)_{n=1}^{N}$, to fit to Equation B.2 in order to estimate ks for virtual sorghum canopies. Input files and scripts to reproduce virtual sorghum canopies used here are accessible on GitHub (see Supplemental Information).

B.1.3 Field experimental sorghum canopies.

To collect incident light data at different depths in the sorghum canopies, two lifts for PAR meters were used to take simultaneous readings in two plots of sorghum. Each of the two plots contained four rows of the same genotype, where the two genotypes varied in their leaf inclination angle. Two sorghum recombinant inbred lines, RIL 63 and RIL 73, were identified as RILs that had varying inclination angles under genetic regulation when QTL mapping was performed in a RIL population generated from a biparental cross of the energy sorghum lines R07018 and R07020 (Bartek *et al.*, 2012). These two genotypes were then planted in 4 row plots with

row spacing of 0.76 m^2 and planting density of 13.2 plants/m^2 in College Station, TX fields (W 96 20, N 30 37).

The consistent depths at which incident sunlight was captured in the canopies were ensured by using two pulley systems that were individually manned. Figure S3 shows the pulley and the PAR meters, LP-80 from Decagon Devices. The platforms holding the LP-80 were raised and lowered in increments of 1 ft (0.3048 m) that were marked on the pulley string to guide the depth in canopy where n are the depth measurements that the PAR meter is n ft from ground, $n = 1, ..., 14$. The genotypes were different heights at the time of measurement, during their vegetative stage; percentages of depth with respect to height of genotype, hgenotype, in the canopies are used such that $x_n = 100\frac{n}{h_{\text{genotype}}}$. The rates of extinction for large and small leaf angle canopies remained the same whether or not depth was scaled, so depth was scaled to remove complexity. One set of readings represents the depths measured from top to bottom or bottom to top by the operators. To mitigate operational biases, the operators, PAR meters, and pulley systems were often swapped between plots. Lastly, consistency of external light conditions between the plots were maintained by using the two PAR meters and coordination between the operators so that the same depths were assayed simultaneously in both plots. Two of the sets of data, those taken on July 25th, were taken after the plots had been thinned to a lesser density; the change in plot density increased the amount of PAR available at lower levels in the canopy, but did not change the relative trend of light extinction (Table B.1).

The platform and pulley mechanism was designed to reduce interference with the 0.865 m probe that contains 80 PAR sensors and was used in between the two middle rows of the 4 row plots. The instrument was set to log PAR data, $I(t)$, automatically at 1 minute intervals, and 3 minutes were spent at each level in the canopy and recorded to ensure that at least two data points were captured at each

level in the canopy. It took users manning the instruments different amounts of time to move the platform to the next level, and so data for the 3 minutes at each level began when both platforms were at the prescribed depth and cloud cover was absent. Light interception in the two plots was simultaneously assayed to ensure that the amount of available sunlight, I_0, was similar. Light intercepted at a given depth was taken as the average of logged PAR measurements that were available in the 3 minute windows. Let the 3 minute interval at depth x_n be $(t_0, t_T)_n$, then $I(x_n) = average_{t\in(t_0,t_T)_n} I(t)$. All plot and plant measurements along with raw depth from the experimental field data and subsequent scaling employed and their fits to Equation B.2 are available in Files SI.

Table B.1: **Light distribution in canopies of RIL 63 and RIL 73.** k is calculated with and without scaling depth (height) and scaled depth; the qualitative relationship of ks between RIL 63 (small angles) and RIL 73 (large angles) remain the same. Data was retrieved in July 2014 on the given dates in College Station, TX."*" denote datasets where RIL 63 was measured by LP-80 #2 and RIL 73 data was measured by LP-80 #1. The LP-80s were switched at other times.

	Height		**Scaled depth**	
Time of acquisition	RIL 73 (large)	RIL 63 (small)	RIL 73 (large)	RIL 63 (small)
July 22nd at 1400*	0.641993	0.426493	0.065483	0.054250
July 22nd at 1530*	1.245529	0.392925	0.127043	0.050097
July 23rd at 1440	0.576724	0.295480	0.060556	0.038412
July 24th at 1240	0.325723	0.332027	0.034201	0.043828
July 24th at 1350	0.389599	0.297924	0.040908	0.040220
July 25th at 1220*	0.374110	0.131663	0.038159	0.017116
July 25th at 1400*	0.395254	0.210151	0.040316	0.031041

B.2 Files SI

B.2.1 Additional supporting folders and files

Additional Supporting Folders and Files is available on `https://github.com/mulletlab/leafangle_supplement/`.

For a detailed work-through of virtual sorghum analysis, custom scripts, .lpy files, and associated parameters are in folder v_sorghum. Data pertaining to field experiments of RIL 63 and RIL 73, and analysis to find ks are in folder exp_fields. Specific parameters and additional information for the PCR amplification of *dw3* or *Dw3* are in folder *dw3*. Genotypes, phenotypes, correlation of phenotypes, genetic linkage maps, heritability calculations and statistics, and multiple QTL mapping penalized LOD scores for each phenotype and the statistics on QTL models presented are in folder h2_and_qtl.

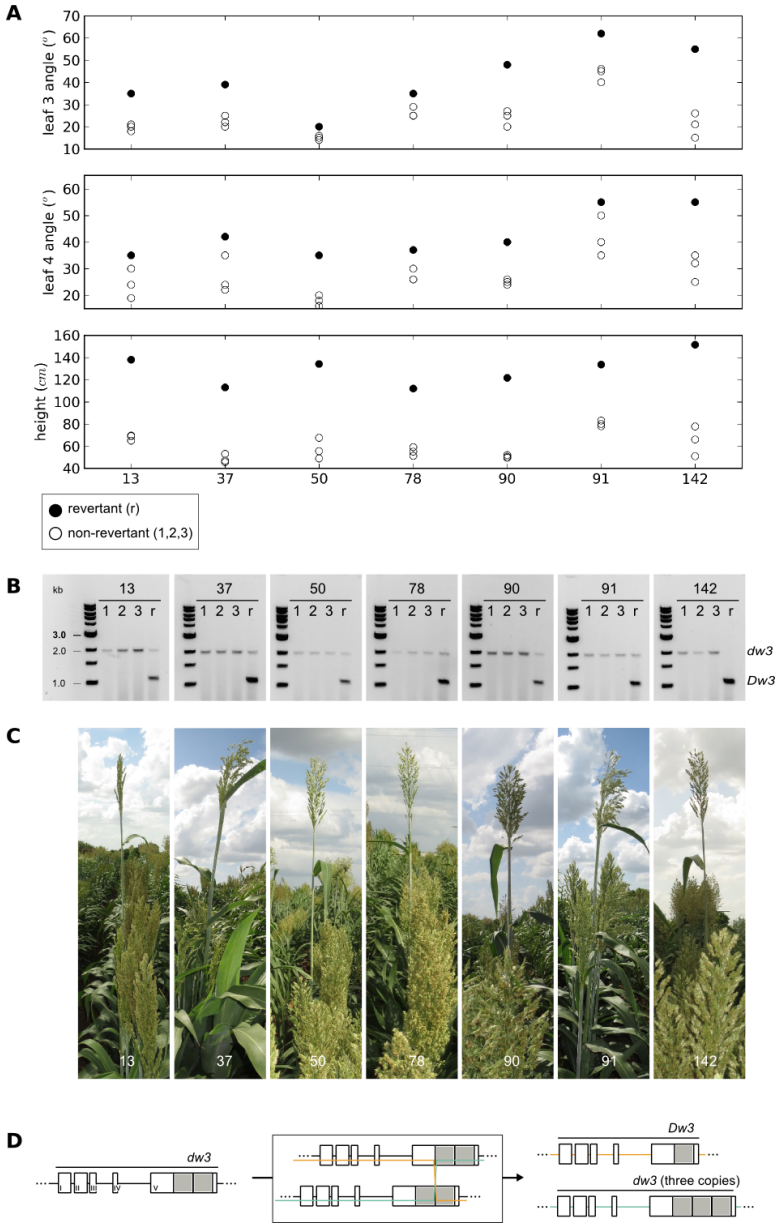

Figure B.1: ***dw3* regulates leaf inclination angle. (A)** Phenotypes of *Dw3* revertants (r; filled circles) and non-revertants (1, 2, 3; open circles) of RILs of BTx623 x IS3620C (B) Genotypes of revertants and non-revertants of RILs at the *dw3* locus generated using primers designed by (Barrero Farfan *et al.*, 2012) that flank the 882 bp tandem repeat that makes *dw3* non-functional. Unequal crossing over at *dw3* causes reversion. (C) Tall revertants in the row with short non-revertants were identified and genotyped (A and B) in the fields in College Station, TX. (D) This diagram shows the *dw3* non-functional allele with tandem repeats on exon V as two gray boxes. Unequal crossing over during meiosis results in a revertant functional *Dw3* and a non-functional *dw3* with three copies of the 882 bp sequence.

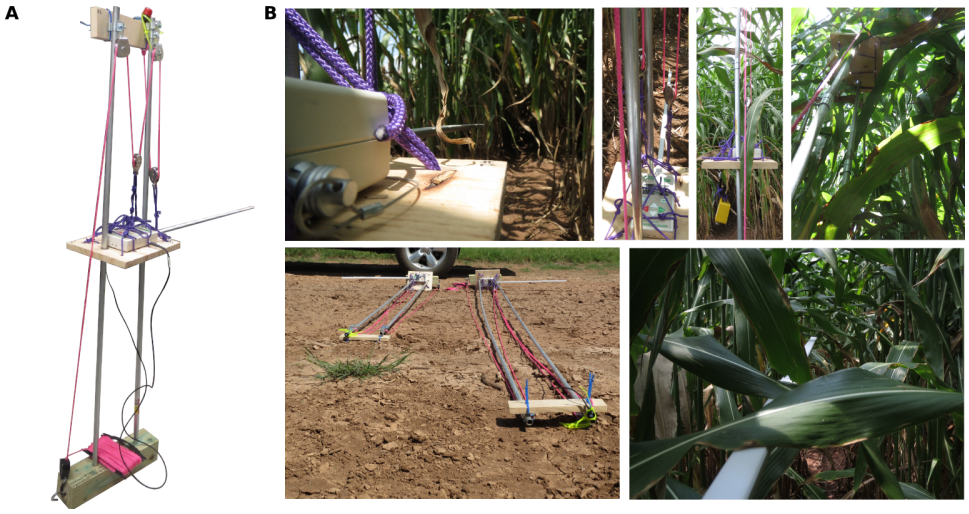

Figure B.2: **Lift for LP-80 PAR meter. (A)** Image of the PAR meter lift used for collection of light measurements within canopies. **(B)** Images of the lift and PAR meter in preparation and in the sorghum canopy. There are views of the wooden platform that the LP-80 sits on and the PAR meter that takes measurements in the canopy. There is also an image of both lifts, where the one on the left has its metal poles extended to 10 ft and the one the right has poles extended to 15 ft.

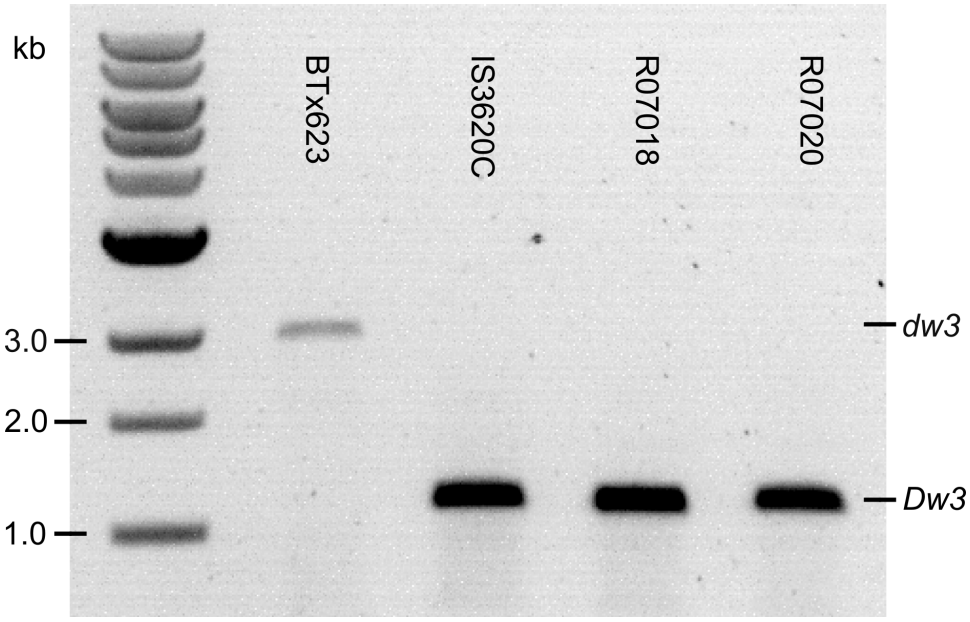

Figure B.3: ***dwarf-3* genotypes of the bi-parental mapping populations.** BTx623, IS3620C, R07018, and R07020 were genotyped at the *dw3* locus generated using primers designed by (Barrero Farfan *et al.*, 2012) that flank the 882 bp tandem repeat that makes *dw3* non-functional. BTx623 has non-functional *dw3* alleles, while IS3620C, R07018, and R07020 have functional *Dw3* alleles.

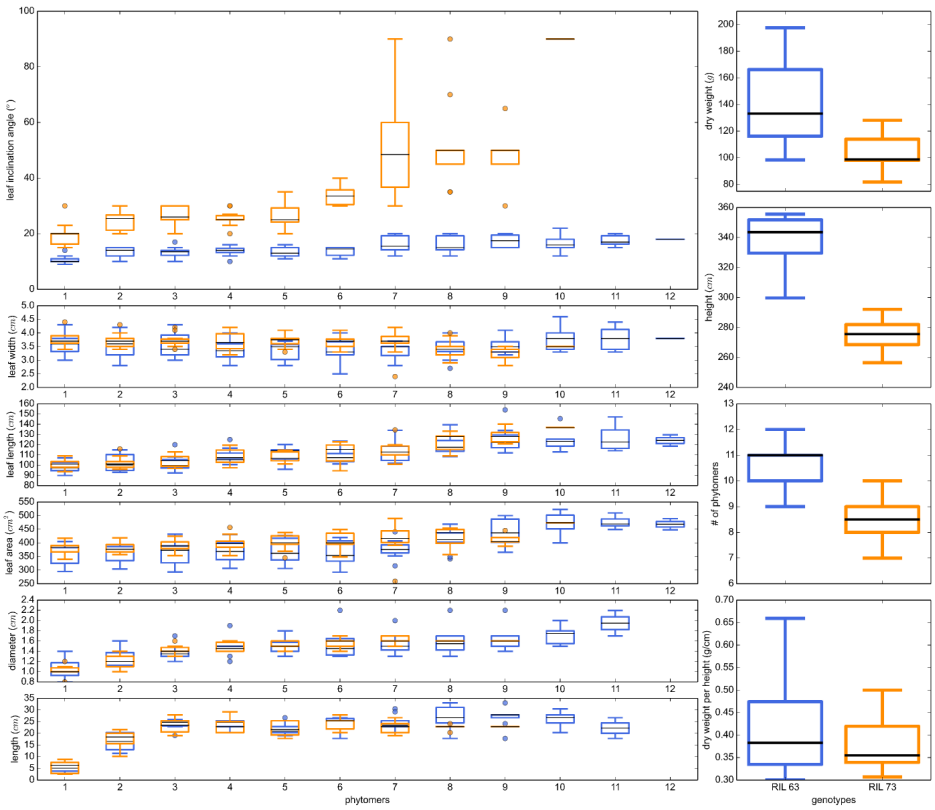

Figure B.4: **R07018 x R07020 RIL 63 & RIL 73 phenotypes.** The boxplots show phenotype data from ten sorghum plants of RIL 63 (blue) and RIL 73 (orange) that were grown in 4-row plots in College Station, TX fields. On the left are measurements by phytomer, where phytomer 1 is at the top of the plant and corresponds to the phytomer with the most recent fully expanded leaf. On the right are measurements of the two genotypes. Leaf chlorophyll (SPAD) was also characterized: RIL 63 with 41.3±.5 nmol/cm^2 and RIL 73 with 44.7±4.0 nmol/cm^2.

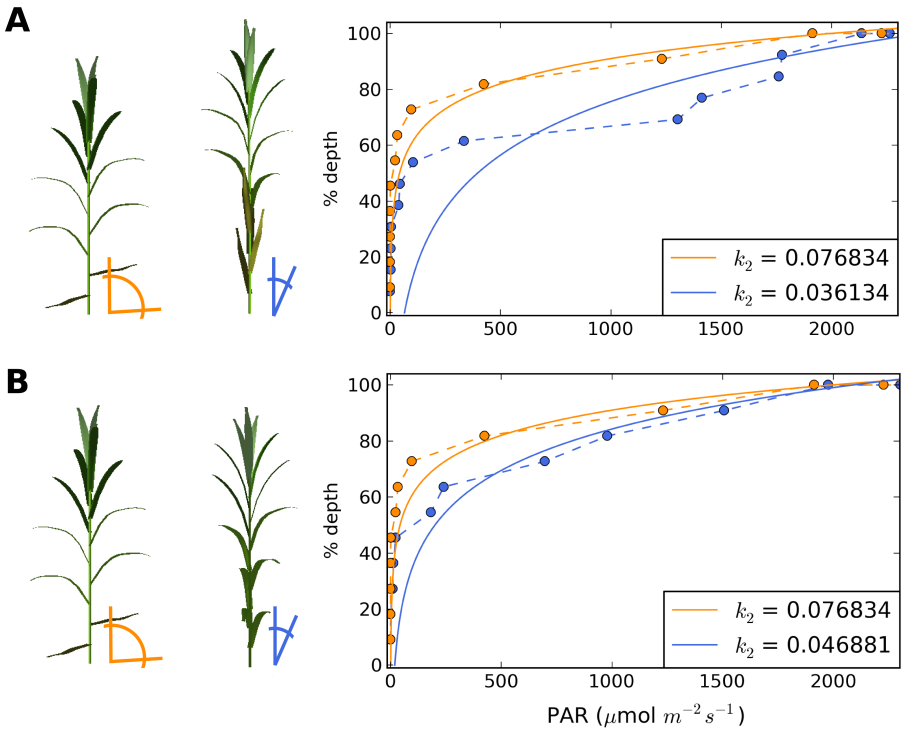

Figure B.5: **Leaf inclination angle regulates light distribution in canopies.** (A) Images of RIL 63 and RIL 73 bioenergy sorghum plants (same as in Figure 2 of the main text). (B) Reducing the height of RIL 63 to be the same height as RIL 73 was done by decreasing the number of phytomers. The resulting k of RIL 63 remains smaller than RIL 73 in solar conditions representative of College Station, TX on July 22, 2014 at 15:30. While the qualitative relationship between k values of these plants remains unchanged with respect to height, the change in k that occurs when shortening a plant with RIL 63 angles indicates that height also plays a role in light distribution.

APPENDIX C

SUPPLEMENTARY MATERIAL FOR BIOENERGY SORGHUM CROP MODEL PREDICTS VPD-LIMITED TRANSPIRATION TRAITS ENHANCE BIOMASS YIELD IN WATER-LIMITED ENVIRONMENTS[1]

C.1 VPD-limited transpiration in simulated water-environments

The VPD-limited transpiration trait model was evaluated in simulated terminal-drought and well-watered environments for its duration in the vegetative stage in order to examine the trade-offs dependent on water in this relatively long developmental stage. Energy sorghum canopy closure occurs ∼75 days after emergence and remains closed until harvest when the crop has sufficient water supply to maintain an LAI ¿ 4 (∼200 days; Figure 4.1 and Figure C.3). Therefore, modeling was initially used to observe the effects of the five m_2 values on biomass accumulation during the summer in College Station, Texas, at a developmental stage where all genotypes have closed canopies (LAI $>$ 4). To aid examination of the impact of variation in m_2, the VPD-limited transpiration trait was induced after canopies had fully developed under fully irrigated conditions so that all genotypes modeled were equivalent until the point of trait induction. After the trait was induced, no additional water was provided to simulate a terminal-drought condition and to observe differences in how the genotypes utilized water in the soil profile in the absence of additional irrigation or rainfall (Figure 4.4C-E). This analysis shows that the rate of biomass accumulation per day declines as m_2 decreases and the total biomass accumulated increases due to increased transpiration efficiency, resulting in an increase in water use effi-

[1]Submitted manuscript "Bioenergy sorghum crop model predicts VPD-limited transpiration traits enhance biomass yield in water-limited environments" by Truong, SK, McCormick, RF, and Mullet, JE (2017) Copyright ©2017

ciency (Figure 4.6C, E). This trade-off occurs because the VPD-limited transpiration trait shifts plant water utilization and biomass accumulation to times of day with lower VPD where transpiration efficiency, $\frac{TE_c}{vpd(t)}$, is higher. The simulated genotype lacking a vpd_{BP} accumulated biomass the most rapidly and used 90% of the water available for biomass accumulation within ∼13 days. In contrast, the genotype with the most restricted transpiration rate at VPD higher than the breakpoint accumulated biomass most slowly, using 90% of the available water in the soil profile in ∼44 days. The restriction of water uptake and use to times of the day when transpiration efficiency was higher resulted in more biomass accumulated under water-limited conditions. A comparison of genotypes with the most extreme m_2 parameters evaluated (m_2=-12.49, 20.26) illustrates this response. Plants where m_2=-12.49 take ∼30 days longer to utilize available water in the soil profile, accumulate an additional 1.235 kg m^{-2}, approximately 32% more than plants lacking a vpd_{BP}, m_2=20.26, which accumulate an additional 0.937 kg m^{-2} after irrigation stops (Figure 4.4C). The additional gain associated with an m_2=-12.49 during the simulated terminal-drought resulted in an overall biomass increase of 18% relative to the plants lacking a vpd_{BP}. Moreover, leaf senescence occurs in sorghum when water limiting conditions induce water stress, which physically reduces transpirational surface area (Blum, 2004). The model includes a water deficit induced leaf senescence response, therefore decreases in LAI occur most rapidly in in genotypes with large m_2 (Figure 4.6D). In contrast to the terminal-drought condition, the VPD-limited transpiration trait has a negative impact on biomass accumulation under water sufficient conditions (Figure 4.6F-H). To assess this potential loss of biomass, the impact of the VPD-limited transpiration trait was modeled by activating the response following canopy closure as described above in the simulated terminal-drought, except in in a well-watered condition. The results show that under well-watered conditions, genotypes that have

the VPD-limited transpiration trait accumulated less biomass, and this impact was greatest on genotypes with the most negative breakpoints (i.e. m_2=-12.49) (Figure 4.6F-H). These results indicate that utilization of the VPD-limited transpiration trait in breeding and deployment of genotypes that express the trait need to be optimized for target environments, and modeling represents a means to identify these optima for further testing.

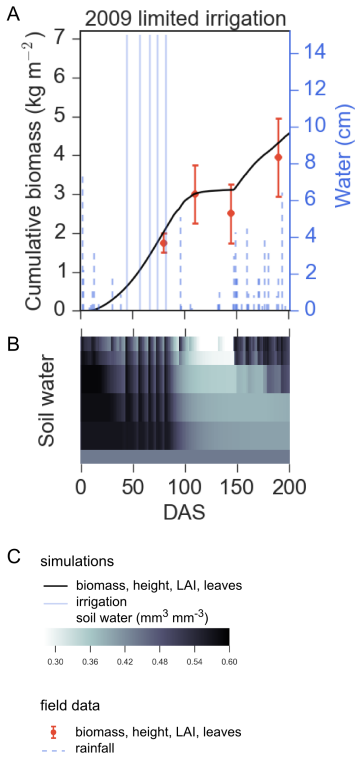

Figure C.1: **Growth simulations of energy sorghum agree with field data for 2009 under the limited irrigation regime.** (A) Predicted biomass accumulation given 2009 environmental conditions under limited irrigation regime falls within the margin of measurement error by the end of the season, (B) Simulated water profile discretized as multiple soil layers. Plotting descriptors are provided in panel C, and experimental data are plotted showing means and error bars representing one standard deviation where available (Olson *et al.*, 2012).

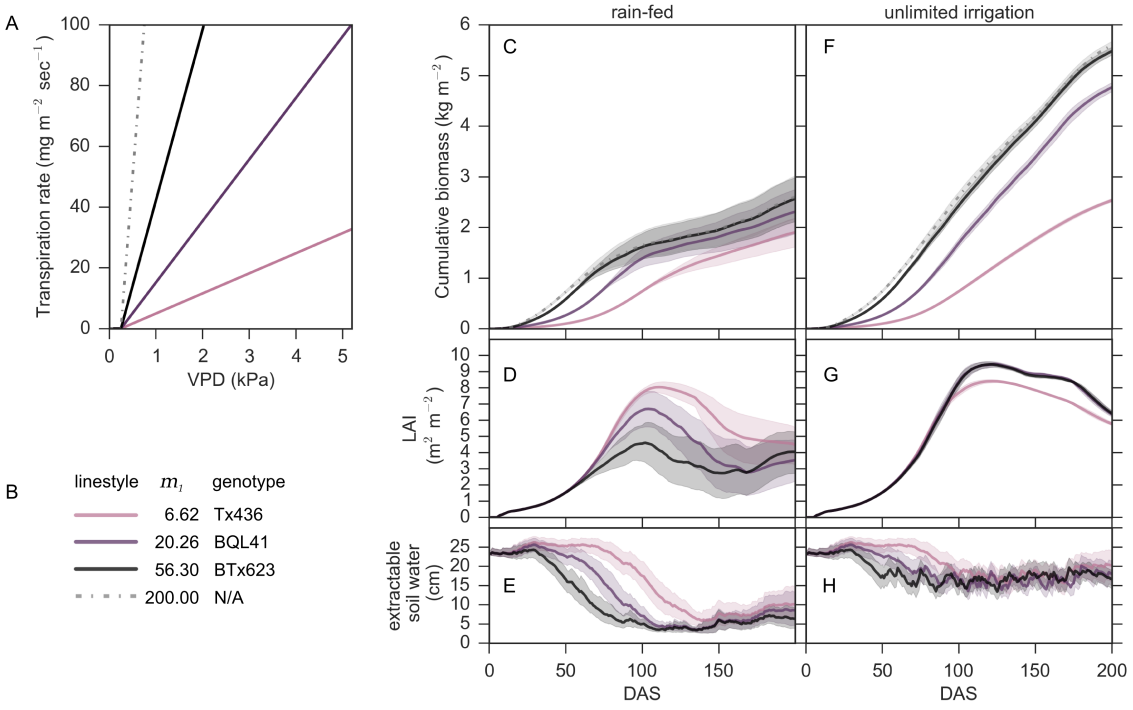

Figure C.2: **Energy sorghum biomass accumulation from 2000-2014 cropping seasons simulated in College Station, TX, with water supply regimes of unlimited irrigation and only rainfed.** (A-D) Growth simulation of plants with no vpd_{BP} ($m_1 = m_2$) are plotted with continuous colored lines that correspond to their $m_1 \in \{6.62, 20.26, 56.3, 200\}$. (C, F) Cumulative biomass, (D, G) leaf area index (LAI), and (E, H) extractable soil water are plotted as the means and 98% confidence intervals estimated with 50,000 bootstraps for the respective rain-fed and unlimited irrigation conditions.

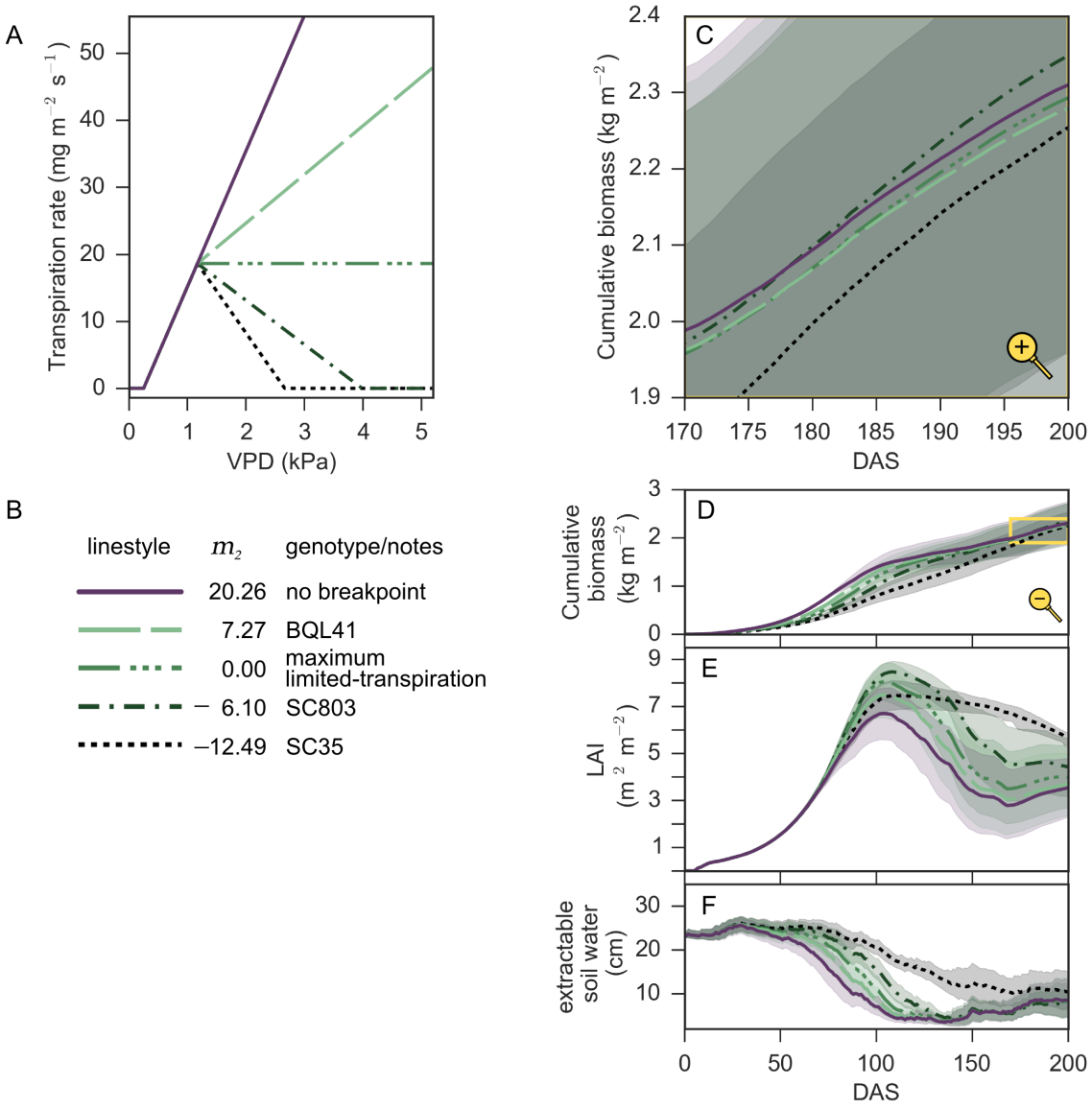

Figure C.3: **Predictions of biomass accumulation by bioenergy sorghum with VPD-limited transpiration traits grown in rain fed cropping conditions of College Station, TX, 2000 2014 against days after sowing (DAS).** (A, B) Diagram of the parameters of VPD-limited transpiration, with variation of transpiration rate after VPD breakpoint, vpd_{BP}, of 1.17 kPA (representative of the BQL41 genotype); $m_2 \in \{20.26, 7.27, 0, -6.1, -12.49\}$. (C-G) The simulations start with a fully saturated water profile at DAS 0. Light bands are 98% confidence interval bands from 50,000 bootstraps. (C) Zoomed in biomass accumulation for DAS 190 200 of panel D. Simulations of (D) biomass (E) LAI and (F) extractable soil water are plotted against DAS 0 200.

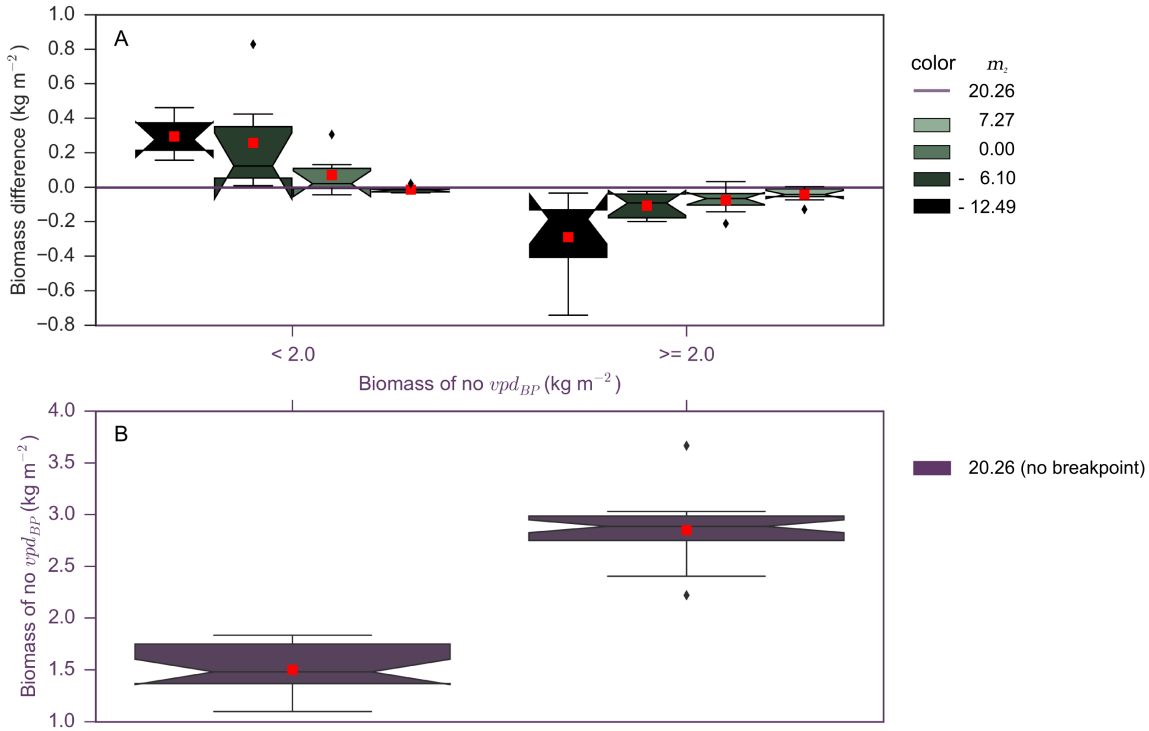

Figure C.4: **Performance of VPD-limited transpiration in low and high yielding rain-fed cropping environments.** (A) Boxplots show the biomass difference between genotypes with a VPD-limited transpiration trait and the genotype that lacks a vpd_{BP} (m_2=20.26). The notches represent the 95% confidence interval of the median, and the red boxes are the respective group means. These statistics are plotted against end biomass accumulated (DAS 200) by the genotype that lacks a vpd_{BP}, where the low and high yielding environments are defined as above and below 2.0 kg m^{-2}. (B) Boxplots of average biomass yields for plants that lack a vpd_{BP} in low and high yielding environments.

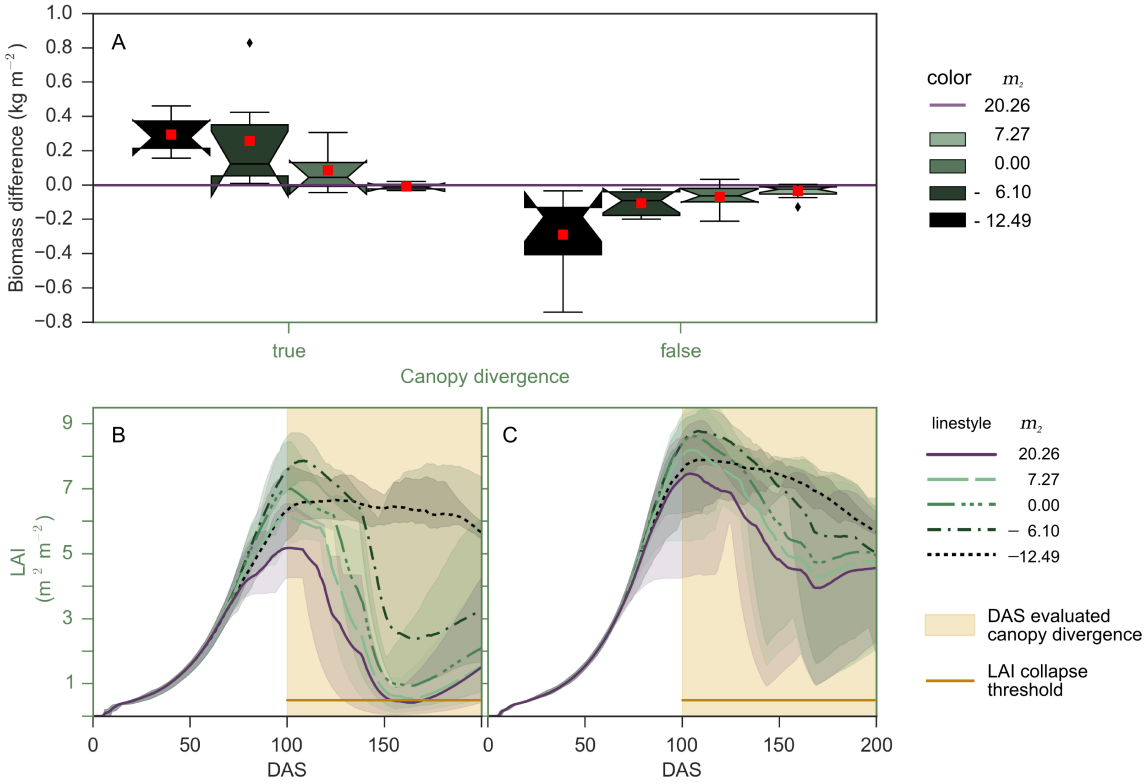

Figure C.5: **Performance of VPD-limited transpiration determined by canopy maintenance.** (A) Boxplots show the biomass difference between genotypes with a VPD-limited transpiration trait and the genotype that lacks a vpdBP (m_2=20.26). The notches represent the 95% confidence interval of the median, and the red boxes are the respective group means. These statistics are plotted against end biomass accumulated (DAS 200) by the genotype that lacks a vpd_{BP}, where the vertical sets of plots are separated by the condition that in the latter half of the cropping season. For each year, for each pairwise comparison between a genotype with and without a vpd_{BP}, and for each day in the latter half of the cropping season ($\geq$100 DAS), if the canopy of genotype lacking a vpd_{BP} senesced (¡0.5 LAI) and the genotype with a vpd_{BP} maintained a larger canopy, then VPD-limited transpiration genotype (with a vpd_{BP}) was predicted to outperform the genotype lacking a vpd_{BP}. (B) LAI is plotted as averages of the conditions met and the more transparent range contains 75% of the data.

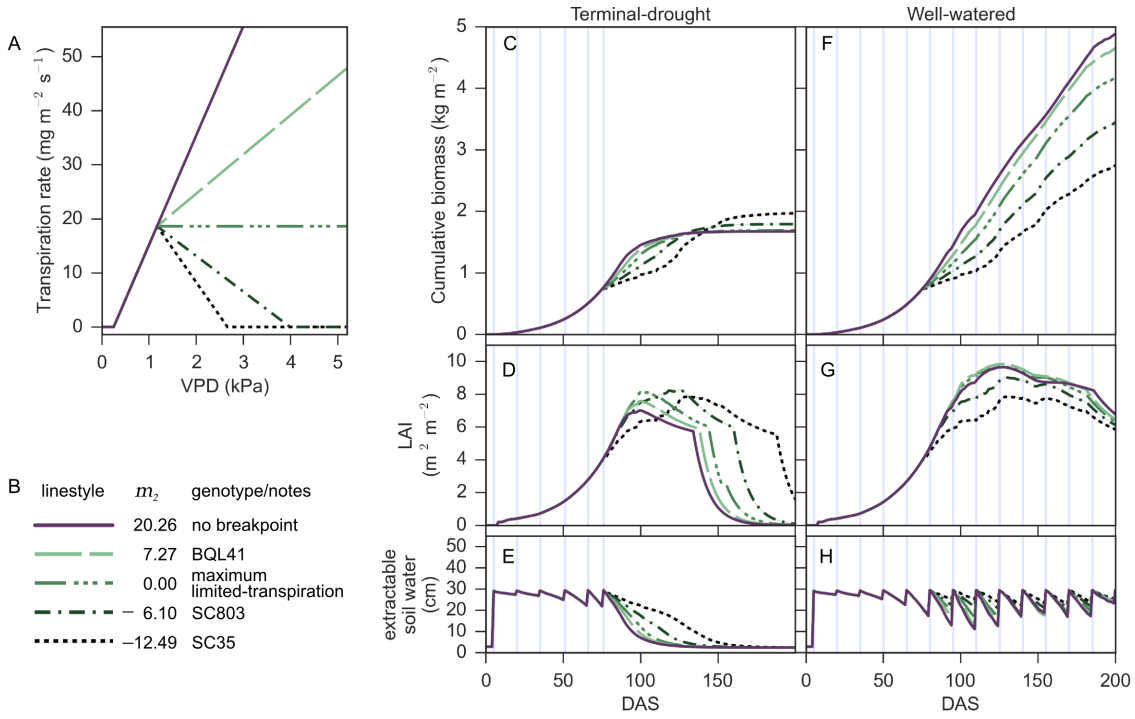

Figure C.6: **Predictions of biomass accumulation by bioenergy sorghum with VPD-limited transpiration traits grown in simulated terminal drought and well-watered environments.** (A) Diagram of the parameters of VPD-limited transpiration, with (B) variation of transpiration rate response to VPD after VPD breakpoint, vpd_{BP}, of 1.17 kPA (representative of the BQL41 genotype); $m_2 \in \{20.26, 7.27, 0, -6.1, -12.49\}$. Simulations of crop physiology where the VPD-limited transpiration trait is induced on day 75 (after sowing; day 0 after canopy closure, LAI ¿ 4) with a fully saturated water profile. The divergence of the VPD-limited transpiration plants illustrates the rate and efficiency of water use of the different responses to VPD. (C) Biomass accumulated, (D) leaf area index (LAI), and (E) extractable soil water in simulated terminal-drought after canopy closure. (F) Biomass accumulated, (G) LAI, and (H) extractable soil water in simulated well-watered conditions.

Table C.1: **Environmental data and end biomass of energy sorghum that vary in their m_2 simulations in the rain-fed cropping environments of College Station, TX.** The table gives data for all years evaluated (2000 2014), as well as further separation to low-yielding (2000, 2005, 2009, 2011 2013) and high-yielding (2001 2004, 2006 2008, 2010, 2014) years that are determined by the drought condition in Figure 5. The VPD-limited transpiration parameters that are fixed here reflect that of BQL41 genotype (m_1=20.26; vpd_{BP}=1.17).

Years (yield)	End biomass of no vpd_{BP} (kg m^{-2})	Daily VPD (kPa)	Yearly rain (cm)	m_2 (mg H_2O m^{-2} s^{-1} kPa^{-1})	Difference in end biomass: m_2 - no vpd_{BP} (kg m^{-2})	(%)
2000-2014	2.31 ± 0.74	3.68 ± 0.41	56 ± 18	7.27	- 0.03±0.04	- 1.34
				0	- 0.02±0.12	- 0.74
				- 6.1	+ 0.04±0.26	1.66
				-12.49	- 0.06±0.35	- 2.42
2000, 2005, 2009, 2011-2013	1.50 ± 0.24	4.03 ± 0.34	43 ± 17	7.27	- 0.01±0.02	- 0.96
				0	+ 0.07±0.12	4.72
				- 6.1	+ 0.26±0.29	17.11
				-12.49	+ 0.30±0.12	19.64
2001-2004, 2006-2008, 2010, 2014	2.85 ± 0.39	3.44 ± 0.23	65 ± 13	7.27	- 0.04±0.04	- 1.47
				0	- 0.08±0.07	- 2.65
				- 6.1	- 0.12±0.07	- 3.77
				-12.49	- 0.29±0.24	- 10.17